Genetic diversity and population structure of the endangered marsupial *Sarcophilus harrisii* (Tasmanian devil)

Webb Miller[a,1], Vanessa M. Hayes[b,c,1,2], Aakrosh Ratan[a], Desiree C. Petersen[b,c], Nicola E. Wittekindt[a], Jason Miller[c], Brian Walenz[c], James Knight[d], Ji Qi[a], Fangqing Zhao[a], Qingyu Wang[a], Oscar C. Bedoya-Reina[a], Neerja Katiyar[a], Lynn P. Tomsho[a], Lindsay McClellan Kasson[a], Rae-Anne Hardie[b], Paula Woodbridge[b], Elizabeth A. Tindall[b], Mads Frost Bertelsen[e], Dale Dixon[f], Stephen Pyecroft[g], Kristofer M. Helgen[h], Arthur M. Lesk[a], Thomas H. Pringle[i], Nick Patterson[j], Yu Zhang[a], Alexandre Kreiss[k], Gregory M. Woods[k,l], Menna E. Jones[k], and Stephan C. Schuster[a,1,2]

[a]Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, PA 16802; [b]Children's Cancer Institute Australia and University of New South Wales, Lowy Cancer Research Centre, Randwick, NSW 2031, Australia; [c]The J. Craig Venter Institute, Rockville, MD 20850; [d]454 Life Sciences, Branford, CT 06405; [e]Center for Zoo and Wild Animal Health, Copenhagen Zoo, 2000 Frederiksberg, Denmark; [f]Museum and Art Gallery of the Northern Territory, Darwin 0801, Australia; [g]Department of Primary Industries and Water, Mt. Pleasant Animal Health Laboratories, Kings Meadows, Tasmania 7249, Australia; [h]National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012; [i]The Sperling Foundation, Eugene, OR 97405; [j]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge Center, Cambridge, MA 02142; [k]University of Tasmania, Hobart, TAS 7001, Australia; and [l]Immunology, Menzies Research Institute, Hobart, Tasmania 7000, Australia

Edited* by Luis Herrera Estrella, Center for Research and Advanced Studies, Irapuato, Mexico, and approved May 23, 2011 (received for review February 24, 2011)

The Tasmanian devil (*Sarcophilus harrisii*) is threatened with extinction because of a contagious cancer known as Devil Facial Tumor Disease. The inability to mount an immune response and to reject these tumors might be caused by a lack of genetic diversity within a dwindling population. Here we report a whole-genome analysis of two animals originating from extreme northwest and southeast Tasmania, the maximal geographic spread, together with the genome from a tumor taken from one of them. A 3.3-Gb de novo assembly of the sequence data from two complementary next-generation sequencing platforms was used to identify 1 million polymorphic genomic positions, roughly one-quarter of the number observed between two genetically distant human genomes. Analysis of 14 complete mitochondrial genomes from current and museum specimens, as well as mitochondrial and nuclear SNP markers in 175 animals, suggests that the observed low genetic diversity in today's population preceded the Devil Facial Tumor Disease disease outbreak by at least 100 y. Using a genetically characterized breeding stock based on the genome sequence will enable preservation of the extant genetic diversity in future Tasmanian devil populations.

wildlife conservation | ancient DNA | population genetics | semiconductor sequencing | selective breeding

Global estimates are that 25% of all land mammals are at risk for extinction (1). Endemic Australian mammals are no exception, with 49 currently named on the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (http://www.iucnredlist.org). Carnivorous marsupials provide striking examples of recent extinction and critical population declines. After the loss of the thylacine (*Thylacinus cynocephalus*), also known as the Tasmanian tiger or Tasmanian wolf, in 1936, the Tasmanian devil (*Sarcophilus harrisii*) inherited the title of the world's largest surviving carnivorous marsupial. Confined, in the wild, to the island of Tasmania, it too is under threat of extinction because of a naturally occurring infectious transmissible cancer known as Devil Facial Tumor Disease (DFTD).

First observed in 1996 in the far northeastern corner of the island state of Tasmania, DFTD has resulted in continuing population declines of up to 90% in areas of the longest disease persistence (2, 3). This rapidly metastasizing cancer is transferred physically as an allograft between animals (4), with a 100% mortality rate. It is predicted that in as little as 5 y DFTD will have spread across the entire Tasmanian devil native habitat, making imminent extinction a real possibility (5).

Cloning and sequencing of MHC antigens has suggested that low genetic diversity may be contributing to the devastating success of DFTD (6, 7). Because MHC antigens can be in common between each individual host and the tumor, which initially arose from Schwann cells in a long-deceased individual (8), the host's immune system may be unable to recognize the tumor as "nonself." On the other hand, a recent study demonstrated a functional humoral immune response against horse red blood cells, although cytotoxic T-cell immunity has not been evaluated to date (9).

An extensive effort is underway to maintain a captive population of Tasmanian devils until DFTD has run its course in the wild population, whereupon animals can be returned to the species' original home range. The strategy for selecting animals for the captive population follows traditional conservation principles (10), without the potential benefits of applying contemporary methods for measuring and using actual species diversity. In hopes of helping efforts to conserve this iconic species, we are making available a preliminary assembly of the Tasmanian devil genome, along with data concerning intraspecies diversity, including a large set of SNPs.

Results

To better assess the genetic diversity of the *S. harrisii* population, we have sequenced the nuclear genomes of two individuals. One animal, named Cedric, was an offspring of parents from northwest Tasmania and survived multiple experimental infections with different strains of tumor, although he eventually succumbed. The other animal, a female named Spirit, came from southeastern Tasmania and was close to death from DFTD when captured. Cedric's genome was sequenced to sixfold coverage on the Roche GS FLX platform with Titanium chemistry, as well as an experimental version of the upcoming XL+ chemistry of

Author contributions: W.M., V.M.H., and S.C.S. designed research; M.F.B., D.D., S.P., K.M.H., A.K., G.M.W., and M.E.J. directed field studies and provided samples; W.M., V.M.H., A.R., D.C.P., N.E.W., J.M., B.W., J.K., J.Q., F.Z., Q.W., O.C.B.-R., N.K., L.P.T., L.M.K., R.-A.H., P.W., E.A.T., M.F.B., D.D., S.P., K.M.H., A.M.L., T.H.P., N.P., Y.Z., A.K., G.M.W., M.E.J., and S.C.S. analyzed data; and S.C.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession no. AFEY00000000).

[1]W.M., V.M.H., and S.C.S. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: vhayes@jcvi.org or scs@bx.psu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1102838108/-/DCSupplemental.

Roche/454 Life Sciences, with read lengths ranging up to 800 base pairs. Roche/454 long read pairs (with inserts up to 17 kb) were used for contig assembly and scaffolding. In addition, Cedric was sequenced on an Illumina platform (GA IIx) to 16.7-fold coverage using paired-end sequencing with short inserts (around 300 bp). Spirit was sequenced to twofold on the Roche GS FLX Titanium platform and to 32.2-fold on the Illumina platform. We also sequenced a tumor taken from Spirit to 19.7-fold coverage. The distributions of coverage depths (determined by aligning reads to the assembly described next) are shown in Fig. 1.

As an intermediate step for measuring intraspecies diversity, we created a de novo genome assembly using the CABOG software package (11); the alternative approach of basing the analysis on comparison with a fully sequenced genome was less attractive because *Sarcophilus* is so evolutionarily distant from the available sequenced marsupial genomes [wallaby, opossum (12)] that many of its genomic regions cannot be accurately compared among those species. The assembly took advantage of the four data types: 454 Titanium paired reads, 454 Titanium unpaired reads, 454 XL+ unpaired reads, and Illumina GA IIx reads, and used reads from both Cedric and Spirit (but not the tumor). See Table 1 for summary statistics and *SI Appendix* for assembly details. The total size of the assembly, about 3.3 Gb (billion bases), is slightly larger than the average for mammalian genomes, but this is to be expected given earlier estimations that the *Sarcophilus* genome size "C-value" is 3.63 (13). Although it was not a main goal of the project to evaluate methods for assembling next-generation sequence data, our project provided an opportunity to compare the performance of two of the better current methods in a real-world setting (*SI Appendix*). Our belief is that the field is not sufficiently mature to allow creation of a definitive reference assembly from data like ours. On the other hand, for assessing genetic diversity and providing a catalog of nucleotide variants, the method works well. It is important to note that by design, the draft assembly resulted from sequencing two individuals to yield a haploid sequence with no variant information. In a subsequent step, Illumina reads were mapped to the assembly and SNPs were called based on differences among the reads, rather than a difference between the reads and the assembly; thus, the SNP calls are largely resilient to assembly errors.

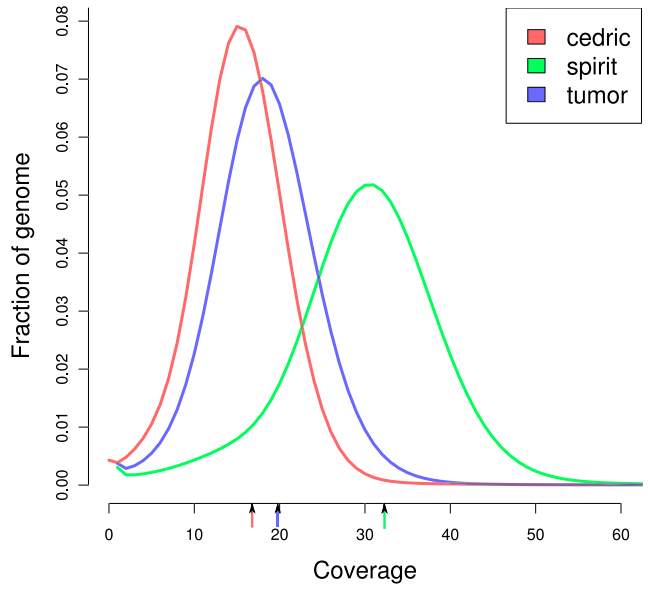


Fig. 1. Sequence coverage depth used for genetic variant detection. The coverage was calculated for Illumina sequences used for our three specimens in SNP calling against a de novo assembled reference sequence (14x coverage 454/Roche and Illumina hybrid assembly), and does not include potential PCR duplicates and secondary alignments. The *y* axis indicates the fraction of the non-N bases in the reference sequence that have a particular coverage. Vertical lines on the *x* axis indicate average coverage for the three samples.

Table 1. Assembly statistics

Contig			Scaffold		
Count	Length (Gbp)	N50 (bp)	Count	Span (Gbp)	N50 (bp)
457,980	2.932	9,495	148,891	3.228	147,544

Mapping the Illumina reads to the assembled contigs let us identify the genetic diversity among the three samples, as well as within each genome (i.e., heterozygosity). We detected 1,057,507 SNPs (i.e., genomic positions where distinct nucleotides can be called with confidence). It is difficult to interpret the SNP count except by comparison with analogous results for species with which we are more familiar. Humans are the only species for which directly comparable data have been published. To avoid effects of methodological differences, we determined SNP counts for several pairs of human individuals exactly as we found Cedric-Spirit differences. Between Cedric and Spirit we found 914,827 substitutions; a southern African Bushman (14) and a Japanese individual (15) contain 4,800,466 SNPs, compared with 3,256,979 for a Chinese individual (16) and the Japanese individual. Surprisingly (given the small number of remaining individuals), lower-coverage Illumina data (5×) indicates that divergence in each of the two threatened orangutan species is about twice that of humans (17).

Classification of nucleotide variants between Cedric and Spirit showed striking differences that indicate a historical mixing of the devil population, in contrast to the ancient separation of the Bushman and Japanese populations or the more recent separation of the Chinese and Japanese populations (Table 2 and *SI Appendix*). In a perfectly mixed population (i.e., matching the hypothesis of "random mating"), there should be twice as many biallelic positions, where both individuals are heterozygous, as where both are homozygous (for different nucleotides). In some sense the departure from the theoretical ratio 2 (see the last row of Table 2) measures stratification between the populations represented by the two individuals. This inference can also be made by considering only heterozygous positions in individuals (Fig. 2*A*) (see *SI Appendix* for details). Although the population subdivision in Tasmanian devils appears to be less deep than that for humans, below we show that a substructure exists and has relevance for efforts to conserve the species.

By sequencing one of five tumors removed from Spirit, we investigated tumor-specific alleles. Using the Galaxy Web site (18) (see *Materials and Methods*), we found 118,575 SNPs that are unique to the tumor: that is, where Cedric and Spirit appear homozygous for the same allele. (By comparison, 198,953 variants are unique to Cedric.) This large number of variants seen only in the tumor confirms that the tumor's source was not a cell from the host, Spirit; rather, the tumor cells contain chromosomes from a different individual. Interestingly, only 20,822 variants were unique to Spirit, which we believe is a result of the presence of Spirit DNA in the tumor sample.

As tumors are likely to contain DNA from both normal and tumor tissue, we estimated the respective amounts by determining the ratio of mitochondrial and nuclear markers that are specific for each. The predicted tumor variants were verified by amplicon sequencing on 110 alleles, thus allowing us to segregate Spirit normal vs. tumor and original host alleles at high sequencing coverage (>1,000-fold). We estimate that 30% of the nuclear DNA and 15% of the mitochondrial DNA in the tumor sample is from Spririt (see *Materials and Methods*). We hypothesize that the difference indicates a higher number of mitochondria per cell in cancerous tissue.

Beside "contamination" from host DNA, there is another inherent limitation to analysis of the tumor sample. Unlike normal/tumor pairings used in other genomic analyses of cancer (e.g., ref. 19), the Tasmanian devil tumors are an infectious cell line, meaning they are "grafted" onto a new host whose genome differs from the original genetic background from which the tumor evolved. Therefore, the genetic analysis must take into

Table 2. Major categories of variant positions between two individuals

Type	Cedric-Spirit	Bushman-Japanese	Chinese-Japanese
SNPs (in millions)	0.91	4.80	3.26
i Heterozygous in both (e.g., AG and AG)	23.8%	10.1%	17.1%
ii Heterozygous in one (e.g., AG and GG)	57.9%	70.5%	68.4%
iii Heterozygous in neither (e.g., AA and GG)	18.3%	19.4%	14.5%
i and *ii*	1.30	0.52	1.18

Minor categories (such as putative triallelic sites) are reported in *SI Appendix*.

account the diploid genome of the present host, the diploid genome of the original host, as well as the somatic mutations of the tumor onto its respective genetic background over many host generations. Although our approach can identify differences between the genomes of Spirit and the tumor, it does not allow us to estimate which of these are somatic mutations that accumulated over time in the tumor cell line. For that identification, it will be necessary to genotype them in a number of individuals so as to identify naturally occurring variants.

We estimate that the number of amino acid differences in the diploid genomes of Spirit and Cedric is roughly 3,000 to 4,000. Although it was outside the scope of this project to predict a definitive *Sarcophilus* gene set, we used the *Monodelphis* genome and its gene annotations to identify 1,141 putative intraspecies protein variants. See *SI Appendix* for more information, including a discussion of how this information might be used to study DFTD.

To estimate the extent and trajectory of *Sarcophilus* genetic diversity since Europeans colonized Tasmania, we sequenced the mitochondrial genomes of seven modern and six historic samples, along with the tumor taken from Spirit. The genomes each contain 16,940 bases of nonrepetitive DNA, together with a short hypervariable region that we did not analyze. The 13 mitochondrial differences between Cedric and Spirit are roughly half the average number for two Europeans and, we estimate, one-sixth the number between two Bushmen (14), an unusually variable human population. Fig. 2*C* compares the number of mitochondrial differences in several species and populations, and indicates that the mitochondrial diversity of *Sarcophilus* is low in absolute terms. On the other hand, the rate that this diversity is decreasing may also be low, as we did not detect much increased diversity in the historic samples (Fig. 2*B*). Excluding the tumor mitochondrial sequence, we detected 24 variable mitochondrial positions. The tumor mitochondria contained an additional five SNPs, but was otherwise identical to that of Spirit, again consistent with the tumor's origin in eastern Tasmania. As the five SNPs from the tumor were not found in the remainder of the population, they may have arisen as a consequence of the increased mutational activity of the tumor tissue.

As our sequencing effort progressed, we were able to construct a series of increasingly extensive genotyping arrays to explore the *Sarcophilus* population structure across Tasmania. We genotyped 17 informative mitochondrial SNPs in 87 wild animals, identifying four persistent mitochondrial haplogroups (denoted A, B, C, and E) (*SI Appendix*, Table S14). Screening an additional 81 wild and 7 captive animals (*SI Appendix*) confirmed region-specific haplogrouping, and identifyied a fifth minor haplogroup, D (Fig. 3*A*). A specimen collected between 1870 and 1910 (OUM5286) showed a unique ancient haplogrouping (denoted hF), but otherwise all of the mitochondrial diversity found in historic samples persists in the extant population.

To provide an opportunity for a higher-resolution analysis of the population structure, we computationally inferred nuclear-genome nucleotide substitutions (20) between Spirit and Cedric as soon as we achieved 0.5× and, later, 2× sequence coverage, generating 96 and 1,536 SNP genome-wide genotyping arrays, respectively. Analysis of 1,532 potential SNP positions identified 702 informative variants used to genotype the 87 wild animals. Using this larger number of SNPs and EIGENSTRAT (21) to draw a principal-components analysis scatter plot (Fig. 3*B*) allows for inferences based on smaller population sizes (in this case, an average of eight per subpopulation) to quantify ancestry. Together with fixation index (F_{ST}) estimates (*SI Appendix*, Table S15) from the 12 geographical locations, nonsex-biased analysis reveals additional subpopulation structure. We note that the plot of Fig. 3*B* roughly recapitulates the geography of the devil samples in a way reminiscent of how human genes have been reported to mirror geography in Europe (22).

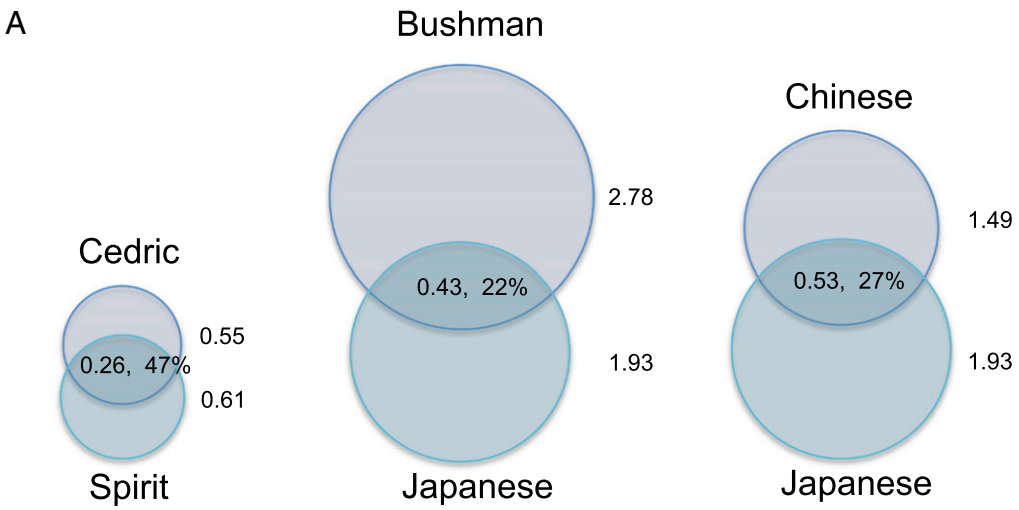
Discussion

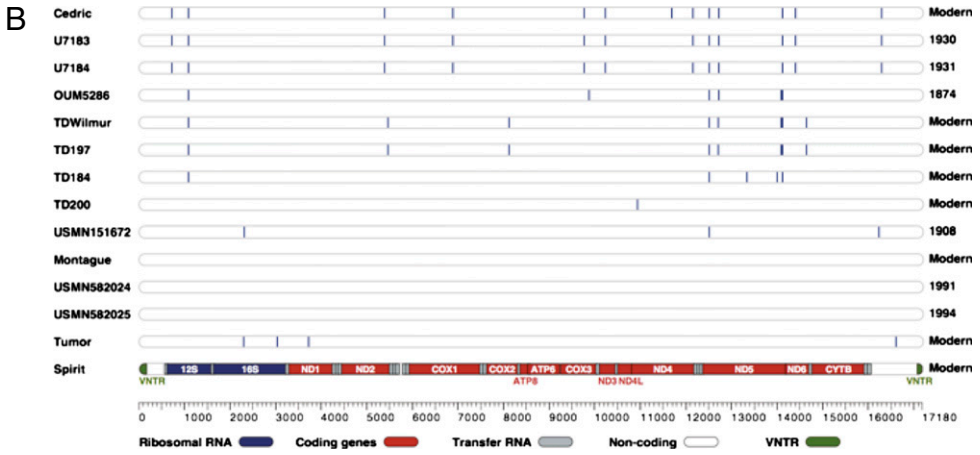
Although most of the capacity of advanced sequencing instruments is currently devoted to resequencing humans (23) and human cancers, interest in sequencing other vertebrates remains alive and well (24). This interest has spawned a growing effort to develop de novo genome-assembly methods that can be applied to data from the so-called next-generation sequencing instruments (25). However, although deep coverage of a vertebrate genome can now be generated in 1 wk on a single instrument, methods for effectively using the data have not kept pace. For example, although the final assembly of the orangutan genome was released in July 2007, the analysis of the data, by a large consortium, was not published until January 2011 (17). Currently, it is not feasible to fully analyze genomes in such depth as quickly as the data can be produced; rather, to keep pace it is necessary to focus the analysis on particular issues. One possibility is to investigate intraspecies diversity, without attempting a definitive analysis of the species' protein sequences.

Although the *Sarcophilus* population is prone to boom-or-bust fluctuations in size (26), the observed near-constancy of mitochondrial diversity over the last 100 y justifies guarded optimism that the species can survive, assuming adequate habitat areas and population numbers and that current diversity can be maintained with the help of a captive breeding program. With the increased sensitivity of using larger numbers of biallelic nuclear markers (vs. only mitochondrial markers), we were able to identify additional population substructure, providing an ideal starting position and rationale for evaluating the on-going breeding program. An alternative to a retrospective analysis of the established breeding population could be random selection of insurance animals guided by the population structure. Our data suggest equal selection from seven zones across Tasmania (Fig. 3*C*), including the diseased region, to ensure adequate capturing of current genetic diversity to supplement and boost current insurance breeding. Indeed, sampling healthy animals in a disease-impacted region may even enrich for alleles offering some protection against DFTD. A third possible use of our data is to genotype a large number of healthy wild animals and select a subset of specified size and sex composition whose overall allele frequencies are as close as possible to a desired distribution; see ref. 27, which also presents a method for optimal selection of ungenotyped individuals from genetically characterized subpopulations (e.g., Fig. 3 *A* and *B*).

Rather than planning a traditional genome-analysis project, our goal is to provide genomic resources to aid conservation efforts for the Tasmanian devil. We are making freely available (*i*) the *Sarcophilus* genomic contigs, (*ii*) alignments of the reads to those contigs, (*iii*) our complete set of 1,057,507 SNP predictions, with allele calls for the three individual samples, and (*iv*) alignments of 121,265 annotated *Monodelphis* protein-coding exons to *Sarcophilus* contigs, covering 17.2 million base pairs, including 1,134 amino acid differences and 1,891 synonymous substitutions among the three *Sarcophilus* genomes (see *Materials and Methods*). Those exons exhibit 91.1% nucleotide identity and 94.7% amino acid identity between *Monodelphis* and *Sarcophilus*, although it should be kept in mind that our procedure strongly favors well-conserved regions.

A potential follow-up study is to search for protein polymorphisms possibly related to an individual's ability to resist or





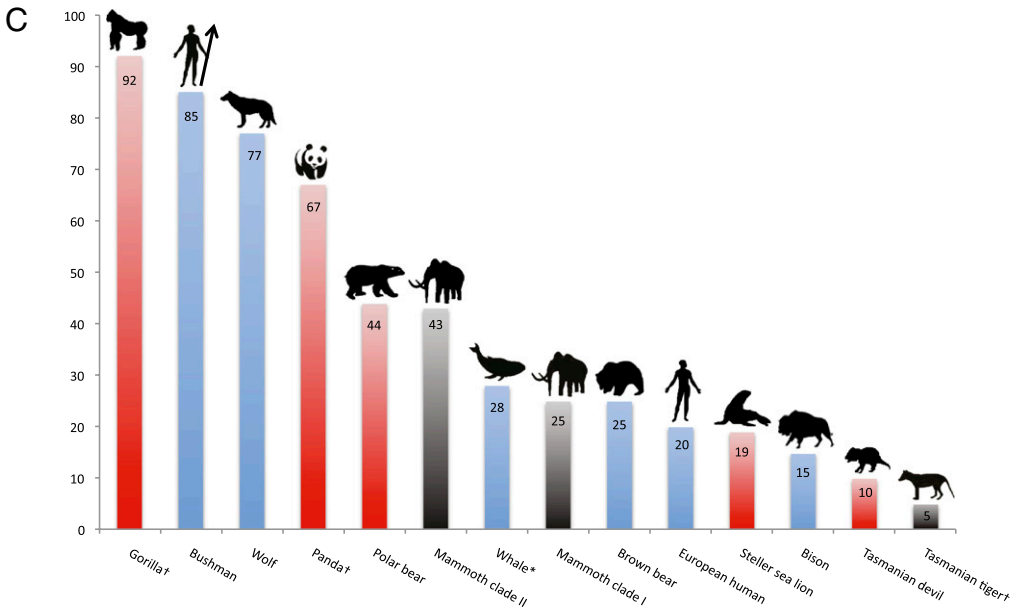


Fig. 2. Genetic diversity of *Sarcophilus*. (*A*) The numbers of heterozygous sites in Cedric and Spirit (in millions), and the number shared between them, compared with two human pairs (the only other vertebrate species for which strictly comparable data are available). *Sarcophilus* has far fewer such sites. In addition, a much higher fraction is shared between individuals, indicating less population stratification than in humans (see *SI Appendix*). (*B*) Mitochondrial diversity covering the last 100 y. Locations of single nucleotide variations (neglecting the hypervariable region) are indicated as vertical lines in the seven modern and six museum specimens relative to the eastern-derived animal, Spirit. Diversity ranges from the geographically most western animal (Cedric) to the most distant eastern animal (Spirit). (*C*) Average numbers of mitochondrial genome differences between pairs of individuals, ignoring hypervariable regions. Species designated by the 2008 IUCN Red List of Threatened Species as "endangered" or "critically endangered" are indicated in red, and extinct species are in black. Species and populations in blue are thriving. †Species represented by only two sequences. *Whales are averaged over five species. Woolly mammoths are divided into two mitochondrial clades (30). The gorillas may be from separate subspecies, *Gorilla gorilla and Gorilla beringei*. It is apparent that mitochondrial diversity is not the only factor affecting species endangerment; habitat loss and other factors are often critical.

delay the onset of DFTD. One speculative case, the *ERN2* gene, is discussed in the *SI Appendix* to illustrate computational methods that can be applied to winnow candidates down in preparation for laboratory experiments. Another line of study, starting with our data, could be to look for differences between the tumor and normal tissues, perhaps using as clues the 138 amino acid variants that we observed only in the tumor (*SI Appendix*, Table S10). In this regard we have validated 110 variants

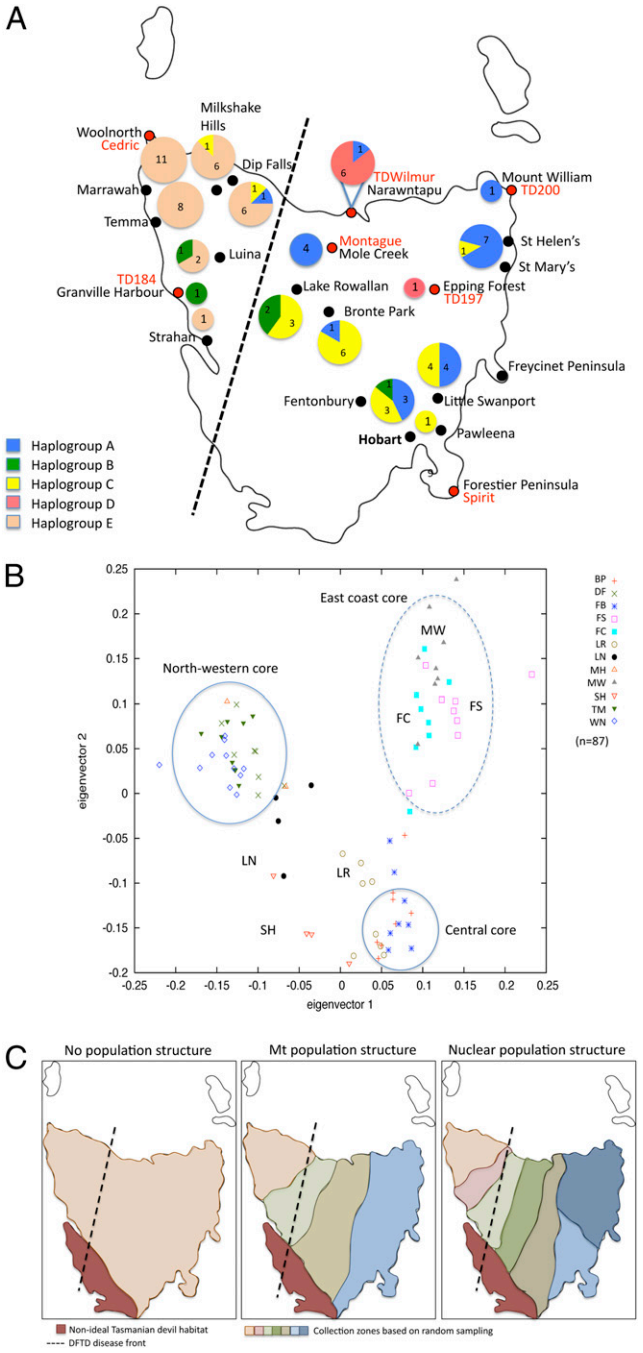


Fig. 3. Population structure of *Sarcophilus*. (*A*) Mitochondrial diversity map of Tasmania for animals from 17 locations, including some now in a mainland captive breeding program. Pie charts depict the location and size (number of animals) of individual populations. Identifiers for modern animals included in the complete mitochondrial sequencing are indicated in red. Four major mitochondrial haplogroups (A, B, C, and E) were identified. A fifth minor haplogroup, D, was found predominantly in the single offshore captive breeding population. (*B*) Principal-components analysis scatter plot for 702 nuclear SNPs genotyped for 87 Tasmanian devils from 12 geographical locations reveals population substructure and diversity. Two core populations of low genetic diversity are found in the northwest and Bronte Park central regions of Tasmania. Although there is clustering of the eastern populations, each adds a unique subpopulation to this broad cluster. Two-letter codes can be inferred from *A*. (*C*) Partitions of Tasmania into regions where equal numbers of individuals for a captive breeding program should be chosen, based on our data.

using amplicon sequencing on a semiconductor sequencing platform (*Materials and Methods*). For example, although we observed no amino acid polymorphisms in orthologs of the well-known cancer-related genes *P53* and *NF1*, *SI Appendix*, Table S10 includes a mutation in the putative *Sarcophilus* gene orthologous to *ANTXR1*, which has been reported to regulate *P53* (28). Although such studies are not the main goal of our work, we have taken a few steps in that direction (*SI Appendix*) to jump-start other efforts, including a preliminary analysis of putative mutations affecting the glycosaminoglycan degradation metabolic pathway. Based on these analyses, we offer three hypotheses that could explain tumor growth and provide initial targets for in-depth immunological and cell biological follow-up studies. We believe that this project illustrates the promise of high-throughput sequencing and genotyping methods for helping to assess intraspecies genetic diversity, including comparison with historical levels, and for planning how to maintain the remaining diversity.

Materials and Methods

Samples for Whole-Genome Sequencing. Two Tasmanian devils were selected to undergo extensive genome-wide shotgun sequencing as references for our SNP-detection approach. Criteria for animal selection included the furthest geographical location of origin as best reflecting current disease spread. Spirit, a 4-y-old female Tasmanian devil, was captured in late 2007 on the Forestier Peninsula (southeastern Tasmania) with severe DFTD. Cedric, a 4-y-old male, was born in captivity to two northwestern parents (maternal line from Woolnorth and paternal line from Arthur River region). Cedric initially demonstrated an antibody response to DFTD, although he succumbed to a later challenge after no further immunization (29). DNA was extracted from whole blood using the Qiagen DNA mini kit (Qiagen Inc.). Samples from Cedric and Spirit (lymphocyte-extracted DNA), were sequenced as described above and assembled as outlined above and described in detail in *SI Appendix*.

SNP Calling. The sequenced Illumina reads were mapped to the CABOG assembly using BWA Version 0.5.8a, allowing up to four differences in reads of length 76/80/82 bp. The reads were soft-trimmed using a "q parameter" of 20 in BWA, ensuring that the low quality bases were not used in mapping. The SNPs were then called using SAMTools Version 0.1.12a. For Cedric (with an average coverage of 15×) SNPs were called in regions, with coverage between 4 and 53, whereas for Spirit (with an average coverage of 29×) SNPs were limited to regions with coverage of 4 to 72. We also filtered to throw-away SNPs with a SNP quality lower than 30. This process enabled us to call 558,270 SNPs for Cedric and 864,664 SNPs for Spirit, compared with the assembled reference. There were 914,827 locations where at least two distinct nucleotides were called for Cedric or Spirit.

We used SAMTools to call the consensus at each individual at each of the variant locations. However, if the coverage at a location was less than six reads, and the consensus call was homozygous for an allele, we identified one of the alleles as "—", which is used to denote insufficient coverage in *SI Appendix*, Table S9. The threshold coverage value (6×) was chosen to reflect the expectation that more than 99.9% of the genome should have a coverage greater than 6 if the average coverage was 15× (the coverage we see in Cedric). To put the comparison of pairs of human genomes on an equal footing, we picked pairs with comparable or higher coverage than Cedric and Spirit, with reads from the same brand of sequencing instrument (Illumina). Data were discarded to reach the levels in the two Tasmanian devils, and the same software pipeline was used to identify homozygous and heterozygous nucleotide differences. The results are shown in Fig. 2*A*. and Table 2.

Ancient Samples. A total of nine museum specimens, in the form of hair shaft collections, were sampled from three museums. Use of hair shafts not only provides a rich source of mitochondrial DNA, it ensures minimal specimen damage during sampling. All samples were stored as dry skins and minimally 20 hair shafts were collected and used for DNA extraction, as previously described (30). 454-sequencing was successfully performed for six specimens, ranging 2 to 100 y before DFTD outbreak. Three specimens from the Smithsonian (Washington, DC) include USMN151672 from 1908, USMN582024 from 1991, and USMN582025 from 1994. The two samples from the Museum and Art Gallery Northern Territory, U7183 and U7184, were stored as complete specimens at room temperature at the museum in Alice Springs, Australia. Both male specimens were trapped in the Welcome and Arthur river regions of northwestern Tasmania in 1930 and 1931, respectively. Specimen

OUM5286 from the Oxford University Museum of Natural History (United Kingdom) was collected between 1870 and 1910 (exact date unknown).

Ethics Approval. Animal collections and sampling were covered by the Animal Ethics Committee of University of Tasmania, Ethics study number 08877 (G.M.W.) and Department of Primary Industry and Water AEC Approval Number 21/2007–08 (M.E.J.).

Genotyping Arrays. Custom designed 96- or 1,536-plexed Golden-Gate SNP array was generated using predicted SNPs selected according to probability of assay success. Genotyping was performed for 87 core samples, including 9 sample replicates, using the Golden-Gate SNP Genotyping assay according to the standard protocol provided by the manufacturer (Illumina Inc.). In brief, assay oligonucleotides were added to a total of 250 ng of genomic DNA for allele-specific extension. The specific extension products were used for the PCRs, followed by purification using 96-well-filter plates. Samples were transferred to a 384-well microplate for hybridization of the Sentrix array matrix chip and the purified PCR products. After washing, the Sentrix array matrix chip was imaged using the Illumina BeadArray Reader (BeadStation 500G) with submicron resolution. Analysis of genotyping data were performed using the Beadstudio software (version 3.2.32) from Illumina (www.illumina.com). The procedure for selecting individuals to genotype is described in the *SI Appendix*.

Tumor Variant Validation. The 112 SNPs that were only called in the tumor were experimentally validated by amplicon sequencing using the Ion Torrent semiconductor sequencing platform. Of the variants, 110 were successfully genotyped in Cedric, Spirit, and the tumor. For Cedric and Spirit there was concordance between the Illumina data and the genotypes except for three cases of apparent mispriming. We confirmed the tumor alleles in 89 instances, often with more than 1,000 reads per variant. Sequencing on the semiconductor platform was conducted according the manufactures manual. These data were also used to estimate the fraction of DNA in the tumor sample that came from Spirit, as follows. For each SNP that was confirmed to be homozygous in Spirit and heterozygous in the tumor, we divided the number of tumor reads with the Spirit allele by the number of tumor reads with the other allele; the average of these ratios was 1.88. Suppose that the fraction x of the tumor sample is from Spirit. Let A be the allele in Spirit and B be the other allele in the tumor. Then the ratio of As to Bs in the sample is $(2x + 1 - x)/(1 - x)$. Setting that ratio to 1.88 and solving for x gives $x = 0.88/2.88 = 0.306$. Thus, this approach estimates that 30% of the nuclear DNA in the tumor sample is from Spirit. (To estimate the analogous figure for mitochondrial DNA, we looked for the Spirit allele at the positions of unique tumor variants.)

Population Structure. We used STRUCTURE (31) v2.2 to determine the population structures of 87 Tasmanian devils for the pilot minimal coverage data. All 69 SNPs, including the linked SNPs, were used. We ran STRUCTURE using the default setting for population number K from 2 to 8. For each population number, we obtained results from five independent runs. Population groupings were based on the average log likelihoods of data and its variance. For phase-two analysis using the extensive number of 921 SNPs, we used the EIGENSTRAT method (21), which identifies population substructure through principal components analysis. Using larger numbers of SNPs and the EIGENSTRAT method allows for inferences based on smaller population sizes to quantify ancestry within samples.

Data Availability. This Whole Genome Shotgun project has been deposited at DDBF/EMBL/GenBank (accession no. AFEY00000000). The version described in this article is the first version, no. AFEY01000000. Alignments of the reads to the assembly can be viewed at http://main.genome-browser.bx.psu.edu. A table containing 1,057,507 putative SNPs is available at the Galaxy server (http://usegalaxy.org), including a variety of information about each SNP, such as the number of reads for each allele in each of Spirit, Cedric, and the tumor, quality values for the SNP calls, and information related to using the SNP in genotyping assays. Alignments of putative *Sarcophilus* coding regions with the *Monodelphis* genome can be fetched by gene name from http://tasmaniandevil.psu.edu and viewed at http://main.genome-browser.bx.psu.edu. Galaxy also provides a table of 3,069 putative SNPs in protein-coding region (1,134 identified amino acid differences).

ACKNOWLEDGMENTS. We thank Erin Noonan, Willow Farmer, Shelly Lachish, and Paul Humphrey for assistance with sample processing; Rodrigo Hamede, Shelly Lachish, Clare Hawkins, Fiona Hume, Billie Lazenby, Dydee Mann, Chrissie Pukk, Jim Richley, Shaun Thurstans, and Jason Wiersma for field collections; Tim Faulkner and John Weigel for captive animal contributions; Gavin Dally for facilitating access to museum collections; Wee Siang Teo for technical assistance; Tim T. Harkins for advice on sequencing technologies; Somasekar Seshagiri for advice on tumor alleles; Bill Murphy for providing *SI Appendix*, Fig. S3; and George Perry for insightful comments about the manuscript. This project was funded by grants from the Gordon and Betty Moore Foundation (to S.C.S. and V.M.H.) for genome sequencing costs, and GENEWORKS, Australia (to V.M.H.) for tumor-transcript sequencing costs; genotyping costs were covered by a donation from the Allco Foundation, Sydney (to V.M.H.); and National Institute of General Medical Sciences (National Institutes of Health) Grant R01-GM077117 (to J.M. and B.W.). V.M.H. is a Cancer Institute of New South Wales Fellow, Australia and S.C.S. is supported by the Gordon and Betty Moore Foundation. Additional support was provided by Roche (to S.C.S.).

1. Schipper J, et al. (2008) The status of the world's land and marine mammals: Diversity, threat, and knowledge. *Science* 322:225–230.
2. Lachish S, Jones M, McCallum H (2007) The impact of disease on the survival and population growth rate of the Tasmanian devil. *J Anim Ecol* 76:926–936.
3. Jones ME, et al. (2008) Life-history change in disease-ravaged Tasmanian devil populations. *Proc Natl Acad Sci USA* 105:10023–10027.
4. Pearse AM, Swift K (2006) Allograft theory: Transmission of devil facial-tumour disease. *Nature* 439:549.
5. McCallum H, et al. (2007) Distribution and impacts of Tasmanian devil facial tumour disease. *EcoHealth* 4:318–325.
6. Siddle HV, et al. (2007) Transmission of a fatal clonal tumor by biting occurs due to depleted MHC diversity in a threatened carnivorous marsupial. *Proc Natl Acad Sci USA* 104:16221–16226.
7. Siddle HV, Marzec J, Cheng Y, Jones M, Belov K (2010) MHC gene copy number variation in Tasmanian devils: Implications for the spread of a contagious cancer. *Proc Biol Sci* 277:2001–2006.
8. Murchison EP, et al. (2010) The Tasmanian devil transcriptome reveals Schwann cell origins of a clonally transmissible cancer. *Science* 327(5961):84–87.
9. Kreiss A, Wells B, Woods GM (2009) The humoral immune response of the Tasmanian devil (*Sarcophilus harrisii*) against horse red blood cells. *Vet Immunol Immunopathol* 130(1-2):135–137.
10. Frankham R, et al. (2002) *Introduction to Conservation Genetics* (Cambridge University Press, Cambridge).
11. Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24:2818–2824.
12. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.
13. Martin PG, Hayman DL (1967) Quantitative comparisons between the karyotypes of Australian marsupials from three different superfamilies. *Chromosoma* 20:290–310.
14. Schuster SC, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943–947.
15. Fujimoto A, et al. (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat Genet* 42: 931–936.
16. Wang J, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456(7218):60–65.
17. Locke DP, et al. (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529–533.
18. Goecks J, Nekrutenko A, Taylor J; Galaxy Team (2010) Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11:R86.
19. Berger MF, et al. (2011) The genomic complexity of primary human prostate cancer. *Nature* 470:214–220.
20. Ratan A, Zhang Y, Hayes VM, Schuster SC, Miller W (2010) Calling SNPs without a reference sequence. *BMC Bioinformatics* 11:130.
21. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
22. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.
23. 1000 Genomes Project Consortium, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
24. Genome 10K Community of Scientists (2009) Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* 100:659–674.
25. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108:1513–1518.
26. Guiler ER (1992) *The Tasmanian Devil* (St. David's Park Publishing, Hobart, Australia).
27. Miller W, Wright SJ, Zhang Y, Schuster SC, Hayes VM (2010) Optimization methods for selecting founder individuals for captive breeding or reintroduction of endangered species. *Pac Symp Biocomput* 15:43–53.
28. Jones GG, Reaper PM, Pettitt AR, Sherrington PD (2004) The ATR-p53 pathway is suppressed in noncycling normal and malignant lymphocytes. *Oncogene* 23: 1911–1921.
29. Kreiss A (2009) The immune responses of the Tasmanian devil and the Devil Facial Tumour Disease. PhD thesis (University of Tasmania, Hobart, Australia).
30. Gilbert MTP, et al. (2008) Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proc Natl Acad Sci USA* 105: 8327–8332.
31. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.