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.

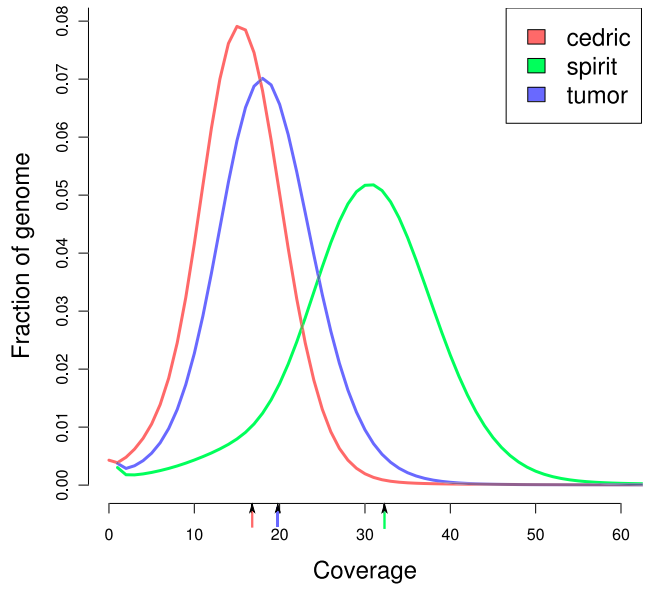

Fig. 1. Sequence coverage depth used for genetic variant detection. The coverage was calculated for Illumina sequences used for our three specimens in SNP calling against a de novo assembled reference sequence (14x coverage 454/Roche and Illumina hybrid assembly), and does not include potential PCR duplicates 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

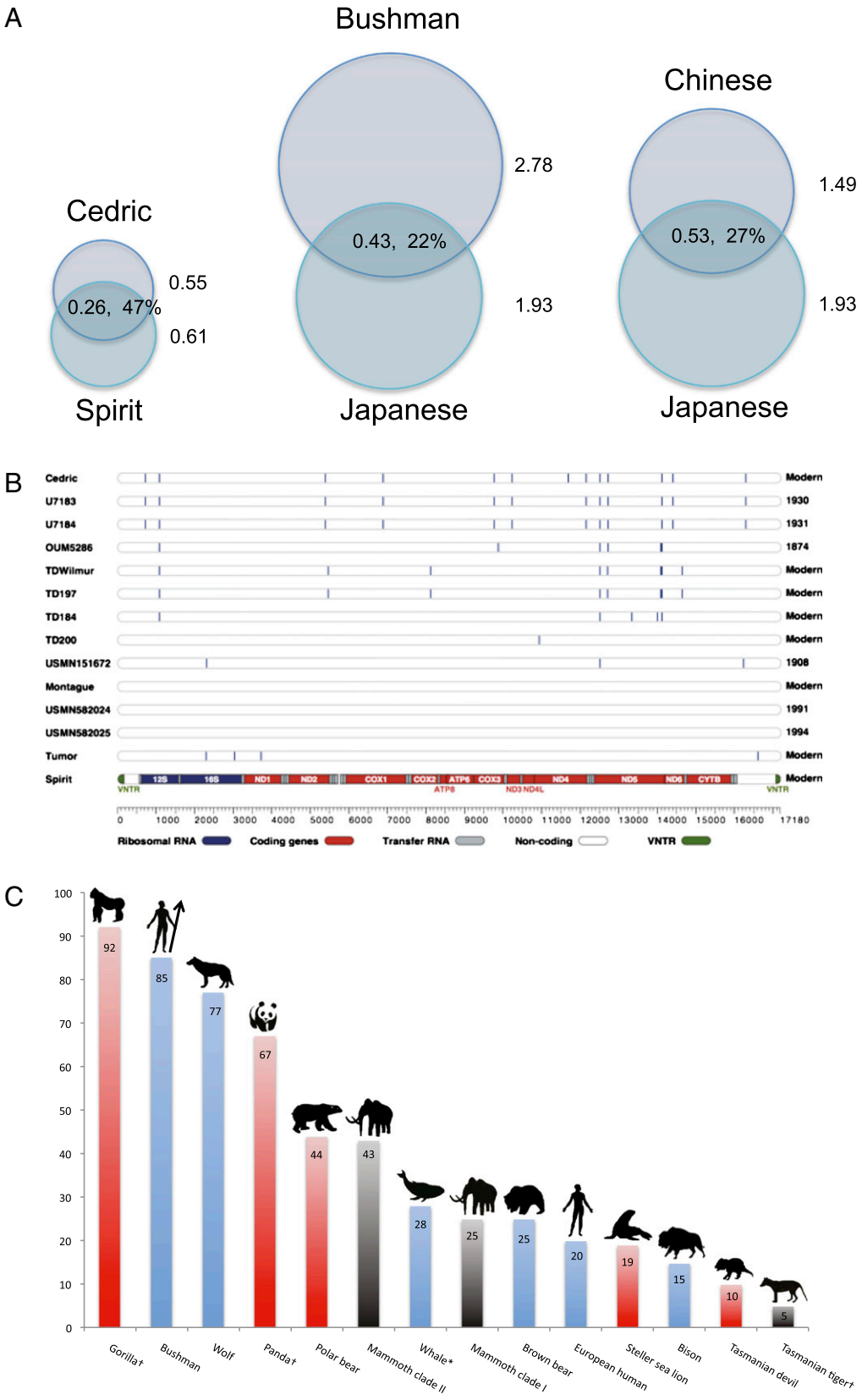

Fig. 2. Genetic diversity of *Sarcophilus*. (*A*) The numbers of heterozygous sites in Cedric and Spirit (in millions), and the number shared between them, compared with two human pairs (the only other vertebrate species for which strictly comparable data are available). *Sarcophilus* has far fewer such 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

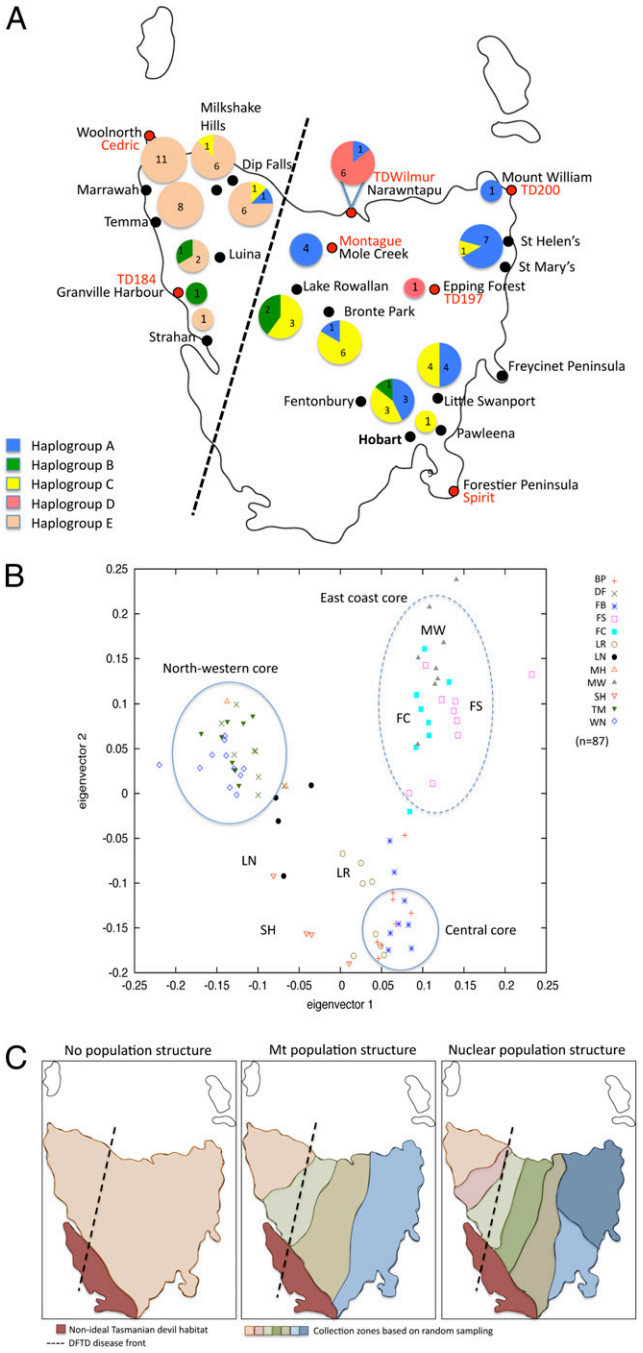

Fig. 3. Population structure of *Sarcophilus*. (*A*) Mitochondrial diversity map of Tasmania for animals from 17 locations, including some now in a mainland captive breeding program. Pie charts depict the location and size (number of animals) of individual populations. Identifiers for modern animals 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.

Supplementary Information for

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

TABLE OF CONTENTS

1. Sequence assembly
   - CABOG Methods.
   - CABOG Results.
   - Newbler Methods.
   - Newbler Results.
   - Assembly Evaluation.
2. SNP calls, classes, and estimated false-positive rate
   - Estimated divergence time from Monodelphis domestica.
   - SNP calls.
   - SNP classification.
   - Random mating.
   - Estimated false-positive rate in SNP calls.
3. Amino-acid differences
   - A potentially functionally relevant variant in the ERN2 protein.
   - ERN2 and human cancer.
4. Mining of metabolic pathways
   - Methods.
   - Results.
     - PRKCH (PKCH)
     - GALNS
     - CCNA-like
     - Other mutations potentially associated with cancer pathology
5. Samples for mitochondrial sequencing, and sequence generation
6. Mitochondrial diversity
   - Data for Figure 1C.
7. Genotyping
   - Captive Sampling.
8. Population structure
   - Notes on Figure 3C.
9. Conclusion
10. References

1. Sequence assembly.

CABOG Methods:

The CABOG assembly was generated with the Celera Assembler (1) software. The pipeline was configured to run the BOG (best overlap graph) module during its unitig stage (2). Source code was downloaded from Source Forge (http://wgs-assembler.sf.net, http://kmer.sf.net) as of 12 October 2010. It was compiled with GNU g++ 4.1.2 on CentOS Linux, kernel version 2.6.18-92.el5. Parameter settings were: merSize = 22, overlapper = ovl, doOBT = 1, unitigger = bog, utgErrorRate = 0.045, utgErrorLimit = 2.5, utgBubblePopping = 1, doExtendClearRanges = 0. All other settings used default values. The high utgErrorRate (default = 0.03) was chosen to exploit LR800 reads presumed to incorporate high error at 3′ ends. The high utgErrorLimit (default = 0) was chosen to incorporate high-error, short overlaps presumed to occur between

some Illumina reads. The low doExtendClearRanges value (default = 2) was selected to reduce run time. Parallel phases ran on a compute grid. Graph phases ran on a 16-core server with 256GB RAM.

The inputs included all the 454 data. Initial processing of the 454 data was performed by CABOG's sffToCA and gatekeeper modules. Processing removed duplicate sequences defined as being identical to the start of some other sequence from the same library. Sequences with high-quality alignment to 454 linker were split into paired reads if each remaining subsequence satisfied the 64bp minimum read length. Other sequences with linker alignments were trimmed of linker and deleted if shorter than 64bp. Assuming a 3Gbp genome, the approximate fragment coverage provided by paired ends was: 1.2X from 15Kbp fragments; 3.9X from 8.3Kbp fragments; 3.3X from 8.0Kbp fragments, and 1.3X from 6.0Kbp fragments. The 15Kbp library sustained the highest losses during initial processing, yielding 1,404,506 reads (495,732 of them paired) from 2,693,262 fragment records in the sequencer's SFF file.

The inputs included a subset of Illumina data: the first read from each pair from all 7 lanes of run 100411. This subset, providing approximately equal coverage to the 454 data, was chosen to balance platform-specific noise in base calls. To avoid overwhelming the limited 454 paired end coverage, the Illumina pairing data was excluded. Initial processing of the Illumina data was performed by CABOG's fastqToCA and gatekeeper modules. Processing trimmed low-quality ends and deleted remaining sequences shorter than 64bp.

Table S1 shows the number of sequences input per library and the number of reads that survived initial processing. It also shows the number of reads that survived CABOG's overlap-based trimming step, which trims each read based on its full set of local alignments to other reads.

CABOG Results:

Table S2 shows high rates of incorporation of reads in contigs. Contigs include 86% to 90% of the reads from every library except one. The 15Kbp paired end library was an outlier in that it contributed 81% of its reads to contigs and 13% to degenerates. (CABOG degenerates are mini-assenblies whose non-repetitiveness was uncertain based on read coverage and whose scaffold placement was precluded by contradictory paired ends.)

Table S3 shows counts of satisfied and violated paired-end constraints in scaffolds. Of the 7,077,508 input reads with a mate pair, 5,327,780 (75%) are co-placed in one scaffold so as to satisfy the constraint.

Table S4 shows selected assembly continuity statistics. The scaffold N50 is 147Kbp. The contig N50 is 11Kbp.

Figure S1 generalizes the N statistic with a cumulative density of bases in unitigs, contigs, and scaffolds. (Unitigs are CABOG's preliminary contigs validated by, but not extended by, the paired end data.) The large gap between the curves for contigs and scaffolds may reflect the lack of short-range paired ends in the assembly input.

The CABOG software incorporates a variant allele recognition algorithm (3) developed for, and tuned on, the human diploid genome (4). The CABOG assembly output includes 1,271,915 variants, each confirmed by high-quality base calls from two or more reads, as described (3). The variants reported in the main paper were derived independently of CABOG's variant calls.

ID	Library Description	Initial Fragments	Processed Reads	Trimmed Reads	Reads with Mate
1	cedric_illumina_100411_1_1	35,277,769	33,176,986	30,305,943	
2	cedric_illumina_100411_2_1	36,344,481	33,803,998	31,029,993	
3	cedric_illumina_100411_3_1	36,165,197	33,638,177	30,881,838	
4	cedric_illumina_100411_4_1	36,409,630	34,000,689	31,224,170	
5	cedric_illumina_100411_5_1	36,271,161	33,757,430	30,999,423	
6	cedric_illumina_100411_7_1	36,237,760	33,791,146	31,014,553	
7	cedric_illumina_100411_8_1	34,023,681	32,027,869	29,442,986	
8	cedric_titanium_15000	2,693,262	2,935,356	1,404,506	495,732
9	cedric_titanium_6000	2,478,692	3,115,570	2,835,768	1,283,132
10	cedric_titanium_8000	3,433,147	4,898,630	3,861,706	2,484,346
11	cedric_titanium_8300	4,890,286	6,306,523	5,557,801	2,814,298
12	cedric_titanium_lr_GB0	4,358,846	4,346,161	3,823,658	
13	cedric_titanium_lr_GB2	4,609,465	4,598,399	4,083,198	
14	cedric_titanium_lr_GBP	5,585,819	5,567,335	4,901,227	
15	cedric_titanium_lr_GBQ	2,651,933	2,641,558	2,326,441	
16	cedric_titanium_lr_GBR	2,789,318	2,779,012	2,460,746	
17	cedric_titanium_lr_GBS	4,204,617	4,187,743	3,686,563	
18	cedric_titanium_lr_GBU	3,727,299	3,713,239	3,285,844	
19	cedric_titanium_lr_GBW	2,947,148	2,938,999	2,566,902	
20	cedric_flx_sr	292,771	264,594	247,495	
21	spirit_flx_sr	465,478	400,708	369,480	
22	cedric_titanium_sr	15,376,177	15,282,463	13,339,946	
23	spirit_titanium_sr	18,829,385	18,759,142	16,470,076	
	CATEGORY SUBTOTALS				
	Illumina 77 bp	250,729,679		214,898,906	
	Titanium paired end reads	13,495,387		13,659,781	
	Titanium long reads	30,874,445		27,134,579	
	FLX standard reads	758,249		616,975	
	Titanium standard reads	34,205,562		29,810,022	

Table S1. Inputs to the CABOG assembly. The library description includes the donor name, the sequencing platform, and an indicator of the library where "lr" indicates LR800 unpaired long reads and "sr" indicates unpaired standard reads. The Illumina input includes 7 lanes of one run but only read #1 of each pair. Processed reads indicate the number of surviving reads after initial processing. Trimmed reads indicate the number of reads that satisfied the 64bp minimum length after application of CABOG's OBT (overlap-based trimming) module. Counts of reads with mate indicate processing to recognize and remove 454 linker. Insert size gives the paired-end distance estimates (in bases, including both read lengths) that were input to CABOG.

ID	Library Description	Fraction Contig	Fraction Singleton	Fraction Degen	Fraction Surrogate
1	cedric_illumina_100411_1_1	87.009%	5.967%	6.276%	0.748%
2	cedric_illumina_100411_2_1	86.821%	6.313%	6.156%	0.709%
3	cedric_illumina_100411_3_1	86.807%	6.328%	6.159%	0.706%
4	cedric_illumina_100411_4_1	86.803%	6.332%	6.159%	0.707%
5	cedric_illumina_100411_5_1	86.811%	6.327%	6.155%	0.707%
6	cedric_illumina_100411_7_1	86.832%	6.316%	6.149%	0.702%
7	cedric_illumina_100411_8_1	86.787%	6.341%	6.162%	0.710%

8	cedric_titanium_15000	81.613%	3.676%	12.941%	1.770%
9	cedric_titanium_6000	87.714%	2.390%	8.537%	1.360%
10	cedric_titanium_8000	87.542%	2.424%	8.669%	1.366%
11	cedric_titanium_8300	88.347%	2.184%	8.140%	1.329%
12	cedric_titanium_lr_GB0	89.658%	0.801%	7.713%	1.829%
13	cedric_titanium_lr_GB2	89.774%	0.782%	7.606%	1.838%
14	cedric_titanium_lr_GBP	89.630%	0.805%	7.737%	1.828%
15	cedric_titanium_lr_GBQ	89.760%	0.784%	7.678%	1.778%
16	cedric_titanium_lr_GBR	89.734%	0.849%	7.563%	1.854%
17	cedric_titanium_lr_GBS	89.419%	0.912%	7.861%	1.807%
18	cedric_titanium_lr_GBU	89.610%	0.881%	7.713%	1.796%
19	cedric_titanium_lr_GBW	89.375%	0.979%	7.855%	1.792%
20	cedric_flx_sr	87.939%	1.804%	8.933%	1.323%
21	spirit_flx_sr	89.032%	1.383%	8.213%	1.372%
22	cedric_titanium_sr	88.376%	1.210%	8.320%	2.094%
23	spirit_titanium_sr	89.861%	0.982%	7.366%	1.791%
	CATEGORY SUBTOTALS				
	Illumina 77 bp	86.838%	6.275%	6.173%	0.713%
	Titanium paired end reads	87.296%	2.448%	8.865%	1.391%
	Titanium long reads	89.621%	0.843%	7.718%	1.818%
	FLX standard reads	88.594%	1.552%	8.502%	1.352%
	Titanium standard reads	89.196%	1.084%	7.793%	1.926%

Table S2. Read fates in the CABOG assembly. Each category of reads shown in Table S1 was measured by its fractional contribution to three categories of assembly output. All quantities are fractions of the trimmed read counts shown in Table S1. Contigs are ungapped assemblies that CABOG deemed non-repetitive and incorporated into a scaffold. Singletons are individual reads that CABOG could not incorporate, presumably due to unique, erroneous, repetitive, or chimeric sequence. Degenerates are ungapped assemblies that CABOG could not incorporate into any scaffold. Surrogates are ungapped assemblies that CABOG treated as collapsed repeats and incorporated at one or more scaffold locations. Reads in surrogates that CABOG localized to exactly one scaffold location are reported as contig reads.

Read Location	Mate Location	Constraint Disposition	Read Count
Scaffold	Same scaffold	Satisfied	5,327,780
Scaffold	Same scaffold	Compressed	34
Scaffold	Same scaffold	Stretched	4,752
Scaffold	Same scaffold	Mis-oriented	1822
Scaffold	Different scaffold	Unsatisfied	636,246
Singleton	Singleton		27,192
Anywhere	Singleton		276,164
Degenerate	Degenerate		273,200
Anywhere	Degenerate		458,550
Surrogate	Surrogate		728
Anywhere	Surrogate		71,040

Table S3. CABOG assembly mate statistics. Given paired end reads, each mate pair is treated as a constraint on assembly. The constraint dispositions are tabulated in categories by read

placement. Satisfaction requires the correct relative orientation and placement within $\mu \pm 3\sigma$, where the per-library mean and standard deviation were derived empirically from the assembly. The table reflects only 454 pair constraints since Illumina pairs were excluded from the input.

Unit	Count	Cumulative	Max	Average	N50
Scaffold (span)	148,891	3,227,539,014	2,942,314	21,677	
Scaffold (bases)	148,891	2,932,162,428	2,818,543	19,693	146,890
Scaffold>2Kbp (span)	67,025	3,113,991,397			
Scaffold gaps	309,089			956	
Contig	457,980	2,932,162,428	135,276	6,402	11,025
Contig>10Kbp	89,517	1,588,254,638	135,276	17,742	
Degenerate	700,328	252,842,858	50,812	361	
Surrogate	25,781	16,823,499	21,071	653	
Surrogate placements	28,761	18,768,868			
Singleton	14,381,809	1,175,915,057			

Table S4. CABOG assembly continuity statistics. Scaffold span includes gap length estimates, whereas scaffold bases does not. Each N50 is based on the cumulative length in that same row. Degenerates are unitigs of 2 or more reads that could not be placed in scaffolds and were therefore excluded from the assembly. Surrogates are unitigs of 2 or more reads that were treated as collapsed repeats and placed at one or more scaffold locations. Singletons are individual reads excluded from the assembly.

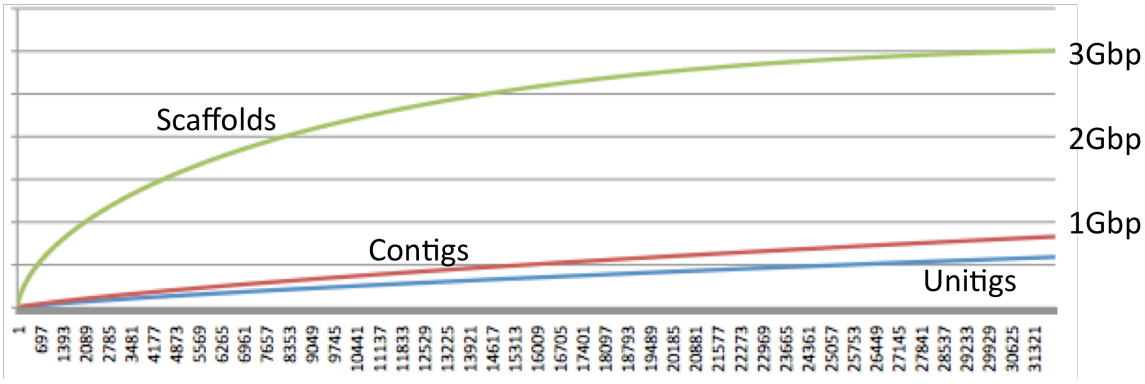

Figure S1. CABOG cumulative continuity. The CABOG assembly was analyzed for cumulative bases in three categories of mini-assembly: unitig, contig, and scaffold. Mini-assemblies are ordered left-to-right from largest to smallest; the largest 32K units are shown for each category. Scaffold size reflects span including gaps between contigs. Colors: green=scaffold, red=contig, blue=unitig.

Newbler Methods:

The Newbler assembly was generated with the R&D 4/19/2010-patch-10/19/2010 version of 454's gsAssembler software (5), an internal R&D version that extends the released v2.5 assembler software. All of the default parameter settings were used, except the use of the –large, -sio and –cpu 24 options, specifying that this is a large genome assembly, that temporary sequential I/O files should be used to speed up the output generation phase, and that 24 processor cores should be used during the parallel phases of the assembly. The assembly was performed on a 256 GB memory, 48 core, shared memory Linux machine. The assembly was completed in 7 days on that system, using half of the processors and with a maximum memory footprint of approximately 216 GB.

The inputs included all of the 454 data plus a subset of Illumina data: both halves of the paired reads from all 7 lanes of run 100411. Standard input processing was applied to the 454 reads. The Illumina reads were quality trimmed using the same algorithm that is used for Sanger reads, which initial tests on other Illumina reads appeared to show satisfactory trimming. Specifically, the 5' and 3' ends of each read are trimmed using a progression of smaller windows (13 bases, 7 bases, 3 bases and 2 bases), testing a read end using a quality score threshold of 20. For the 13bp window, if the 13 bases at the end of the read have an average score less than 20, the trim point is moved in by one. For the other three windows, if the window contains more than 2, 1 or 0 bases with a score less than 20, respectively, then the trim point is moved in by one. In this test, a base of 'N' or 'X' is considered to have a quality score below 20, regardless of the given score. Once a read's end passes the test for a window, the next window is tested. The trimpoint is set to the spot passing the last window test. And, any read trimmed to shorter than 20 bases is discarded. While it is likely that only the 3' trimming was necessary for the Illumina reads, it was easier to pass those reads through the same trimming process.

From that point forward, the standard assembly algorithms handled the combined set of reads, performing alignments, constructing the assembler's internal data structures, and generating contigs and scaffolds. No adjustments were made to handle the Illumina reads, with the exception that the pairing information was not used for Illumina reads. They were treated as unpaired reads in the assembly, and the scaffolding used only the 454 paired-end reads. (The data structure designed for tens of thousands of Sanger paired-end reads was not capable of handling 240 million pairs.)

Table S5 shows the number of input reads and bases for each type of input given to the assembler. It also shows the number of reads and bases of each type that were used in the assembly, after initial processing and trimming. Finally, the table shows the percent of each library aligned in the assembly (reads left out of this percentage are either singletons, chimeric reads or come from repeat regions whose copy-count is too high to remain in the data structures).

Newbler Results:

Table S6 shows the outcomes and computed statistics for each of the 454 paired-end libraries used in the assembly. Note that the computed pair distances, computed separately for each input file, remain consistent across the regions and runs of the same library, and that the number of paired-end reads found in the input, as well as their occurrence in the same scaffold, is relatively consistent across the libraries. The only deviation is a slight decrease in the number of paired-end reads for the 15kb library. (Note: These counts do include duplicate paired-end reads found in the input, and the higher rate of duplicate reads found for the longer library accounts for the difference in paired-end rates here compared to the CABOG metrics. And, while duplicates are counted in these metrics, only single representatives from each set of duplicates is used in the actual scaffolding and consensus calling process.)

Table S7 shows the core assembly statistics. Newbler by default outputs scaffolds whose span is larger than 2kb, which includes both multi-contig scaffolds linked by paired-end reads as well as individual contigs longer than 2kb which could not be unambiguously linked to neighboring contigs. It also outputs "large contigs" consisting of all contigs larger than 500 bp. The scaffold N50 is 657Kbp. The contig N50 is 9Kbp.

Input Read Type	Initial Reads	Initial Bases	Used Reads	Used Bases	% Align
Illumina #1	250,729,679	1,875,586,420	250,425,582	1,323,504,189	88.7%
Illumina #2	250,729,679	1,875,586,420	247,730,187	1,230,497,966	88.7%
Titanium SR	34,205,562	3,134,959,741	34,203,626	3,053,670,729	92.5%
Titanium paired	13,495,387	4,387,961,078	21,297,802	3,923,228,489	89.3%
Titanium LR800	30,874,445	2,492,460,938	30,874,241	2,406,046,145	92.4%
FLX Standard	758,249	183,780,999	758,241	183,372,471	92.7%

Table S5. Inputs to the Newbler assembly. The number of input reads and bases are those of the trimmed reads as they appear in the input file. The Illumina #1 and #2 libraries correspond to the first and second files of the Illumina 100411 run; each file was processed separately so the Illumina data was effectively unpaired. The Titanium SR, LR800, and FLX Standard libraries all provided unpaired reads. The Titanium paired libraries include 4 insert sizes. All of the reads derived from Cedric except portions of the Titanium unpaired and FLX Standard. The breakdown of reads by insert size and by donor are given in Table S1. Used reads and used bases indicate what was used to generate the assembly, after trimming, and after paired-end read separation on 454 data. Each separated paired-end is counted as two used reads. Percent aligned includes the reads that were aligned in the internal assembly data structure used to construct contigs and scaffolds.

Input File	# Paired-End	# in same scaffold	Computed Pair Dist.	Computed Deviation
GOA7F6B01	342,277	187,237	6,398.2	1,599.5
GOA7F6B02	309,842	169,596	6,397.9	1,599.5
GOC1UV301	351,287	190,353	6,553.0	1,638.2
GOC1UV302	328,215	179,121	6,402.0	1,600.5
GOA6XWI01	425,781	237,814	9,440.2	2,360.1
GOA6XWI02	392,692	218,913	9,437.8	2,359.4
GOAXFSH01	382,699	205,971	9,445.0	2,361.3
GOAXFSH02	362,254	195,229	9,459.6	2,364.9
GOC1VQG01	275,675	150,011	9,447.7	2,361.9
GOC1VQG02	201,016	108,182	9,453.3	2,363.3
GOCTLEJ01	441,904	243,802	9,438.2	2,359.6
GOCTLEJ02	411,344	227,564	9,443.6	2,360.9
GLDIVUQ01	562,762	318,012	9,995.0	2,398.7
GLDIVUQ02	516,314	289,319	10,000.7	2,500.7
GLDM0PU01	478,762	282,984	9,986.3	2,496.6
GLDM0PU02	404,723	240,169	9,995.5	2,498.9
GLOQYU002	503,329	297,907	9,994.0	2,498.5
GHGV0J201	334,581	192,693	14,829.9	3,707.5
GHGV0J202	282,086	160,609	14,847.6	3,711.9
GHKH4NR01	265,949	151,986	14,762.7	3,690.7
GHKH4NR02	274,448	155,407	14,791.3	3,697.8

Table S6. Newbler assembly mate statistics. During the scaffolding process, paired-end reads that either occur in the same contig, or are found to occur in the same scaffold, are used to compute the average pair distance and standard deviation for each paired-end input file. The input file names are listed by a run prefix followed by "01" or "02" describing the region of the run.

Unit	Count	Cumulative	Max	Average	N50
Scaffolds	72,995	3,204,626,469	5,310,409	43,901	657,294
Large Contigs	576,580	2,962,902,366	96,023	5,138	8,803

Table S7. Newbler core assembly statistics. Scaffold lengths include gap length estimates. Each N50 is based on the cumulative length in that same row.

Assembly Evaluation:

The LR800 reads were produced by an experimental sequencing protocol. Figure S2 explores base call accuracy in these data. The LR800 reads were sampled and mapped to a preliminary assembly by Newbler. Titanium unpaired reads were sampled and mapped for comparison. The LR800 reads demonstrated higher accuracy up to base 400 but higher error in bases beyond the 400^{th}. Both assemblies appear insensitive to the LR800 error profile as indicated by weak library effect on contig incorporation. See Table S2 and Table S5.

One assembly was to be used as a substrate for read mapping and variant detection. Since the assembly with larger contigs would be most useful for subsequent analysis, the CABOG

assembly was chosen on the basis of its larger contig N50. The CABOG assembly was submit to GenBank. The Newbler assembly is available by request from the authors.

The CABOG assembly was evaluated by comparison to the Newbler assembly, which constituted an independent treatment of the read data. The assemblies were aligned with ATAC software (http://kmer.sf.net). Contig-level alignments compared the CABOG devil.ctg.fasta file and the Newbler 454LargeContigs.fna file. Scaffold-level alignments compared the CABOG devil.scf.fasta file and the Newbler 454Scaffolds.fna file. In the CABOG output, the contig sequence is essentially the scaffold sequence exclusive of gaps. In the Newbler output, the contig and scaffold file content can differ due to different minimum length requirements for inclusion in each.

Table S8 gives a statistical analysis of the alignments. ATAC provides indel-free one-to-one alignments called matches. The cumulative span of ATAC contig matches covered 2.7Gbp or 93% of the CABOG contig span. The cumulative span of ATAC scaffold matches covered 2.6Gbp or 80% of the CABOG scaffold span. ATAC also provides combinations of matches, called runs, constructed to be maximal but distinct. The cumulative span of ATAC contig runs covered 2.8Gbp or 80% of the CABOG contig span. The cumulative span of ATAC scaffold runs covered 2.9Gbp or 89% of the CABOG scaffold span. These ATAC alignments to the Newbler assembly provide confirmation of local sequence arrangement and they span the majority of the CABOG assembly. These results support the utility of the CABOG assembly as a substrate for read mapping.

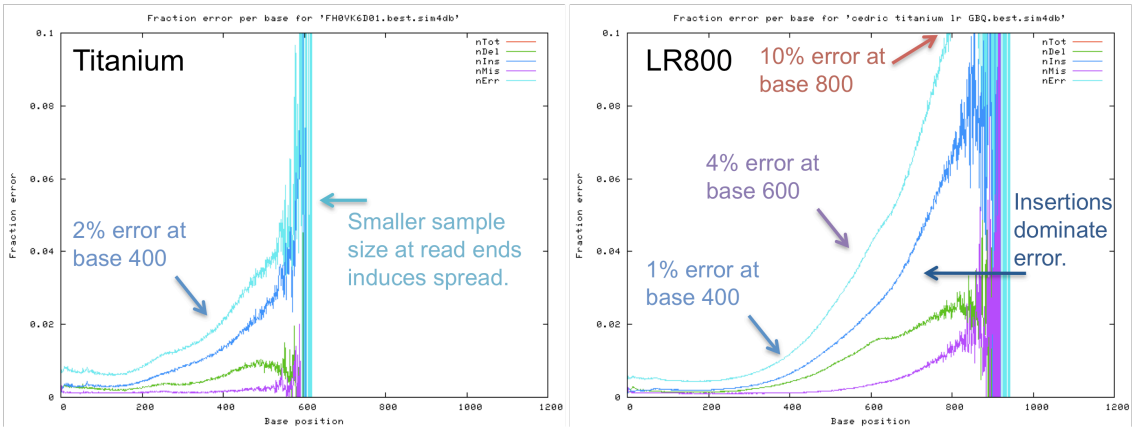

Figure S2. Error profile of reads from selected libraries. Reads were aligned to the contig consensus sequence of an initial Newbler assembly of 454 unpaired reads. Every disagreeing base was scored as an error. For every value *x* on X-axis, the plot shows the fraction of reads of length *x* or more that had an error at *x*. Colors: cyan = total error, violet = substitution error, blue = insertion error, green = deletion error. The figure compares arbitrary subsets of reads from two library types. Left: type = Titanium single read, library = cedric_titanium_sr, file = FH0VK6D01.sff. Right: type = LR800, library = cedric_titanium_lr_GBQ, files = GBQ7AKT[01-08].sff and GBQ8P5V[01-08].sff.

	Count	N50	Cumulative span on CABOG	Cumulative span on Newbler
CONTIG				
Ungapped matches	7,425,828	743	2,717,986,962	2,717,986,962
Maximal runs	653,497	7,157	2,810,238,931	2,816,732,721
SCAFFOLD				
Ungapped matches	7,129,772	758	2,584,636,912	2,584,636,912
Maximal runs	253,469	40,813	2,871,108,992	2,912,932,983

Table S8. Alignment statistics. The Cabog and Newbler contigs were aligned at the contig and scaffold levels. The ATAC software found maximal exact matches that were unique in both assemblies. It found the maximal extensions allowing substitutions but not indels. The table reports statistics for matches of 100bp or more. ATAC also found maximal combinations of matches covering non-overlapping regions of both assemblies. The table reports statistics for runs of any size. The table reports N50 measured on the CABOG assembly based on the CABOG cumulative span. Run lengths differ on each assembly. Runs can span gaps represented by consecutive Ns in scaffolds. By construction, runs on contigs (or scaffolds) cannot span those contig ends (or scaffold ends) whose sequence is repeated within either assembly.

2. SNP calls, classes, and estimated false-positive rate.

Estimated divergence time from Monodelphis. Our strategy for identifying nucleotide substitutions was motivated by our belief that none of the available sequenced genomes could be used as a "reference" for mapping the short sequence fragment generated by our sequencing instruments, based on the following evolutionary analysis. Much of the timetree for select mammalian species (Figure S3), including the Tasmanian devil, was taken from the published literature (6, 7). The divergence time for the thylacine relative to other marsupials was estimated using a bayesian relaxed clock (8, 9), and analyzing published mitochondrial DNA sequence (10). Divergence time estimates for the Tasmanian devil and quoll were based on a relaxed molecular clock analysis of published nuclear (*IRBP* and *FBR*) and mitochondrial (*16S rRNA*) gene sequences.

Figure S3. Composite timetree indicating divergence times based on published estimates (see text) for select mammalian species, and on our comparisons of Tasmanian devil and its nearest cousins the quoll and the thylacine (Tasmanian tiger).

SNP calls. We used SAMTools to call the consensus at each individual at each 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 Table S9, below. The threshold coverage value (6X) 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 15X (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 was 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 Table 2 and Figure 1A of the main paper.

SNP classification. We classified the genomic positions (here called SNPs) where more than one nucleotide could be observed in the sequences from Cedric and Spirit. They fall into 11 categories. At any position, we call zero, one, or two (identical or different) nucleotides in each individual. If two different nucleotides were called in one individual (either Cedric or Spirit), but the other individual had insufficient coverage to call even one nucleotide there, we denote the position's types as (AB,-). Another way to detect differing nucleotides is where both

individual have one called nucleotide, but the calls are different: denoted (A,B). There are three ways to detect differences when two nucleotides are called in one individual but only one in the other, depending on whether the individual with two called nucleotides is homozygous or heterozygous, and in the latter case whether the position is biallelic or tri-allelic. Finally, there are six possible cases when two nucleotides are called in each individual. Fortunately, the number positions with insufficient coverage to call two nucleotides in both individuals, and tri-allelic position (which are probably enriched for sequencing error) is often small enough that it can be ignored. The major categories are: both individuals heterozygous (AB,AB), one individual heterozygous (AA,AB), and neither individual heterozygous (AA,BB). Here we give the full version (Table S9) of Table 2 in the main paper, though now percentages are taken relative to all SNPs.

Table S9. Fractions of different nucleotides seen in pairs of diploid genomes in each of 11 categories. See text, above, for a description of the categories and an interpretation of the results.

Type	Cedric-Spirit	Bushman-Japanese	Chinese-Japanese
(AB,-)	0.071%	0.345%	0.053%
(A,B)	0.473%	0.101%	0.112%
(AA,B)	1.217%	4.654%	1.664%
(AB,A)	2.590%	6.304%	2.307%
(AB,C)	0.004%	0.009%	0.009%
(AA,BB)	15.020%	17.135%	13.829%
(AA,AB)	55.361%	62.417%	65.557%
(AA,BC)	0.005%	0.031%	0.020%
(AB,AB)	25.202%	8.964%	16.424%
(AB,AC)	0.056%	0.040%	0.026%
(AB,CD)	0.000%	0.000%	0.000%

Random mating. Assume random mating, pick an autosomal biallelic position, and let p be the minor allele frequency. Consider two animals from the population. The probability that they are both heterozygous at the position is $4p^2(1-p)^2$, while the probability that they are homozygous (for the respective alleles) is $2p^2(1-p)^2$. Thus the former has twice the expected frequency of the latter, irrespective of p. Thus, considering all autosomal biallelic positions, random mating implies that the expected number double heterozygotes is twice the expected number of double homozygotes.

To take the reasoning somewhat further, let p_1 and p_2 denote the frequency of allele A at a biallelic site (with alleles A/a) in the populations represented by two individuals, respectively. The ratio of the probabilities of observing double heterozygotes versus observing double homozygotes from the two individuals is

$$R = \frac{4p_1(1-p_1)p_2(1-p_2)}{p_1^2(1-p_2)^2 + (1-p_1)^2 p_2^2} = \frac{4\lambda}{1+\lambda^2}$$

where $\lambda = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} \in [0,\infty)$. It can be easily checked that the ratio R takes value in the interval [0,2], and the maximum is achieved at $\lambda = 1$, i.e., when $0<p_1=p_2<1$. That is, the ratio equals to 2 if and only if the two individuals belong to a homogeneous population with random

mating. Otherwise, the ratio is smaller than 2. It is further seen that, as λ deviates away from 1 (either towards 0, which indicates $p_1<p_2$, or towards infinity, which indicates $p_1>p_2$), the ratio R decreases to 0. As a result, we can use R to measure the genetic distance between two individuals. A smaller value of R indicates a larger difference in allele frequencies between two individuals. In practice, we estimate R by the ratio of the total number of double heterozygotes versus double homozygotes. This estimation is only in the average sense because the R ratio may be different at different SNPs. The observations add precision to the informal discussion of Table 2 in the main paper.

For the numbers of heterozygous positions indicated in Figure 1A, we reason as follows. The numbers of heterozygotes within the two devils are similar, but the numbers of heterozygotes within the humans (Bushman versus Japanese, Chinese versus Japanese) are quite different. Let p_{1i} and p_{2i} denote the probabilities of observing a heterozygote at a given SNP i in individual 1 and 2, respectively. We can measure the departure of equal heterozygote frequencies $p_{1i}=p_{2i}$, for all i, as follows: let X and Y denote the number of heterozygotes observed in the two individuals, respectively, and let Z denote the number of double heterozygotes; the measure is given by $T=(X-Y)/\mathrm{sqrt}(X+Y-2Z)$, where $T\sim N(0,1)$ if the two individuals are not stratified ($p_{1i}=p_{2i}$, for all i). It is easily checked that, if $p_{1i}=p_{2i}$, for all i, $E(X+Y-2Z)=\mathrm{Var}(X)+\mathrm{Var}(Y)$, and thus T is a standard test for the difference of means.

The three values are: –75 for devils, 433 for Bushman vs. Japanese, and –286 for Chinese vs. Japanese. These numbers all show significant deviation from a homogenous population, but the extent of deviation (absolute value) is proportional to the mean difference of heterozygosity (p_1-p_2), and to the square root of sample size n, i.e., $E(T) \sim (p_1-p_2)\mathrm{sqrt}(n)$. There are 1 million SNPs used to find heterozygotes in devils, but 4 millions in human. If we treat the data as obtained from independent random sampling schemes, the statistics for humans should be divided by 2, and thus we have –75 for devils, 217 for Bushman vs. Japanese, and –143 for Chinese vs. Japanese. Devils clearly showed less stratification in terms of standardized heterozygosity difference than humans.

Estimated false-positive rate in SNP calls. At an intermediate stage of this project, we called SNPs using roughly 2-fold genome coverage of Roche/454 data and designed a 1536-SNP genotyping array based on those calls. Since many of our current SNP calls, based on roughly 20 times more raw DNA sequence data (Illumina), can be found in the arrayed set, this provided an opportunity to estimate the false-positive rate in our SNP calls. In particular 1133 of the arrayed SNPs are in our current set of over 1 million putative SNPs. Of these, 989 were validated as being polymorphic in Cedric and Spirit (which were included in the set of 87 individuals that were genotyped with the array), though 48 arrayed positions were genotyped as containing the identical nucleotide in all four alleles. (The remaining 96 arrayed positions were not completely genotyped in the two animals.) These observations provide an estimation of the false-positive rate in our SNP calls of 48/(48+898) ~ 5%. However, we are currently genotyping the 48 discordant positions using PCR and Sanger sequencing, which we believe will substantially lower this prediction.

3. Amino acid differences

The number of amino-acid differences between a random pair of humans is roughly 10,000; when one is a Bushman, the number is closer to 14,000 (11). Assuming that the ratio of SNPs to amino-acid differences is roughly the same between the Tasmanian devil and humans, this suggests that there are 3,000-4,000 amino acid differences between two unrelated Tasmanian

devils. To test this hypothesis, we tried the following more direct approach. We selected a longest member of each overlapping set of Ensembl-annotated *Monodelphis domestica* coding regions and aligned it with our *Sarcophilus* scaffolds, requiring at least 75% identity and a unique best match. Under those strict conditions, 13.6% of the annotated *Monodelphis* coding bases aligned to the Tasmanian devil. We used the *Monodelphis* annotations to determine the reading frame, and found that of the SNPs in the aligning regions, 44% of the 1102 SNPs were non-synonymous. This crude estimation gives 3,566 amino-difference between and within (i.e., heterozygous) Cedric and Spirit.

To produce a partial set of putative protein-coding regions, we relied on monDom5 assembly and ENSEMBL gene predictions for *Monodelphis domestica*, downloaded from the UCSC Genome Browser. To account for alternate splice forms, we extracted a pairwise non-overlapping set of 175,692 predicted protein-coding exons. The corresponding nucleotide sequences were aligned to the reference *Sarcophilus* assembly with a program called "lastz". We retained exons where the best alignment (i.e., most nucleotide identities} (1) covered the exon without an insertion of deletion, (2) did not imply a stop codon in the *Sarcophilus* interval, (3) had at least 75% identity in nucleotides and in amino acids, and (4) did not align to a *Sarcophilus* interval that overlapped the region aligned to another *Monodelphis* putative exon. This filtering retained 121,265 exons with 17.2 million basepairs. We detected 1,141 amino-acid differences and 1,895 synonymous substitutions among the three *Sarcophilus* genomes. Between *Monodelphis* and *Sarcophilus* in these putative protein-coding intervals, we observed 91.1% nucleotide identity and 94.7% amino acid identity, though these regions are expected to be depleted of fast-evolving genes, given that they align at high identity despite separation for roughly 100 million years. We applied the SIFT software (12) annotate a marginally less conservative set of 1,167 amino-acid differences with computational predictions of whether the change is "tolerated" (808 cases) or "deleterious" (359 cases).

A potentially functionally relevant variant in the ERN2 protein. To illustrate the kinds of computational analyses that can be applied to predicted intra-species differences of amino acid sequences, we discuss the *Sarcophilus* ortholog of the human *ERN2* gene. We of course realize that we haven't proved an association of this variant and the ability to resist DFTD; out purpose is simply to provide a real-life example of how to begin investigating putative protein variants.

One method that we use to identify potentially interesting amino-acid differences (within or between species) is to look for a high degree of conservation among mammals at that position, which we interpret as suggesting that the amino acid plays a functional role (13). The alignment available at the UCSC Genome Browser (http://genome.ucsc.edu/) provides an easy way to perform this screening. For instance, Figure S4A shows that an amino acid difference that we observed between Cedric and Sprit affects a position that is otherwise invariant among sequenced mammals.

In order to evaluate the possible effects of the R⇒H mutation, we built a model (Figure S4B) of Tasmanian devil ERN2 protein sequence from the known crystal structure of human IRE1 (pdb entry 1hz6), using the SWISS-MODEL server (http://swissmodel.expasy.org//SWISS-MODEL.html). The site of the mutation is near the dimer interface. One reasonable hypothesis to explore is that the arginine makes explicit stabilizing contacts with the other monomer of the dimer. If these contacts are lost by the histidine it might destabilize the dimer. One must however keep in mind that we do not know for sure that the Tasmanian devil protein is a dimer. The only possible candidate for a side-chain that could form a salt bridge with the R is a glutamine at position 635. (It is the E at the center of the sequence YVWQREGLRKV). It is also possible that the H could also make this salt bridge.

In summary, we have looked at the structure and report a single possible feature of interest, a glutamate in the vicinity of our variant. Secondly, we raise the question of whether an arginine at the site of the mutation could form a salt bridge that a histidine could not, although we have no definitive answer to this scenario.

ERN2 and human cancer

Mining of published microarray data using the Oncomine 3.0 Cancer Profiling Database (http://www.oncomine.org/main/index.jsp) showed ERN2 (IRE1b) to be significantly under-expressed in a number of human cancers, including breast cancer (n=336, P=1.6E-7), renal cancer (n=254, P=3E-5) (Data Link: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109), glioblastoma (brain cancer) (n=38, P=1.3E-5) (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7602), hepatocellular carcinoma (liver cancer) (n=35, P=7.5E-5) (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6764) and in both lung cancer (n=72, P=5.8E-8) (https://array.nci.nih.gov/caarray/project/woost-00041) and non-small cell lung cancer (n=42, P=4.9E-6) cell lines (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8332). These studies may suggest a potential tumor suppressor activity of ERN2, although not further studies have been published. A single study has shown over-expression of ERN2 in colon cancer (n=302, P=1E-44) and colorectal cancer (n=71, P=6.2E-6) (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109).

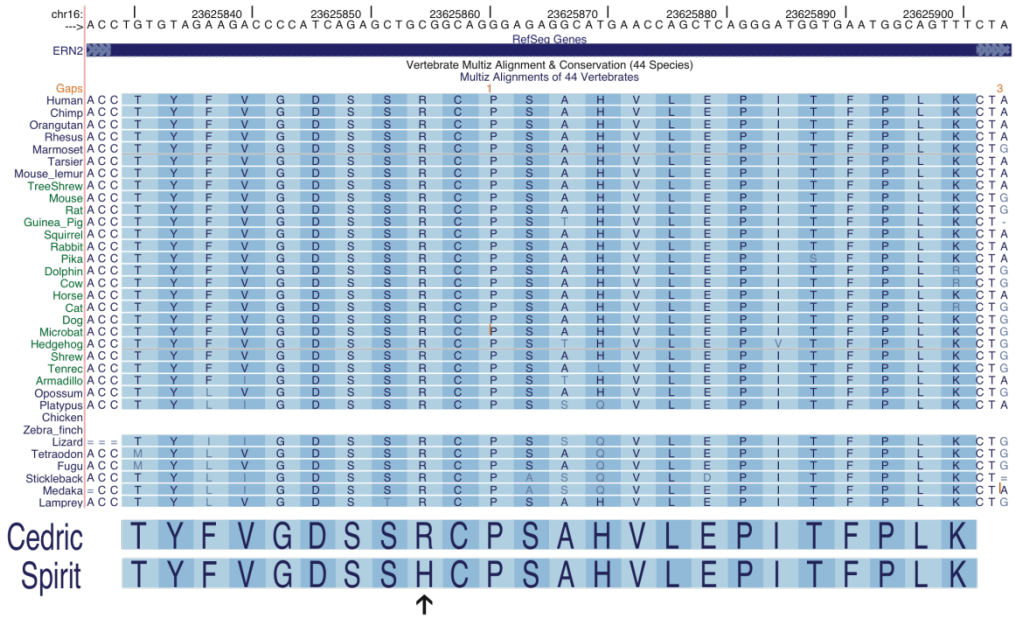

Figure S4A. ERN2 amino acid sequences (written from C-terminus to N-terminus) compared among 36 vertebrates to Tasmanian devil Cedric and Spirit.

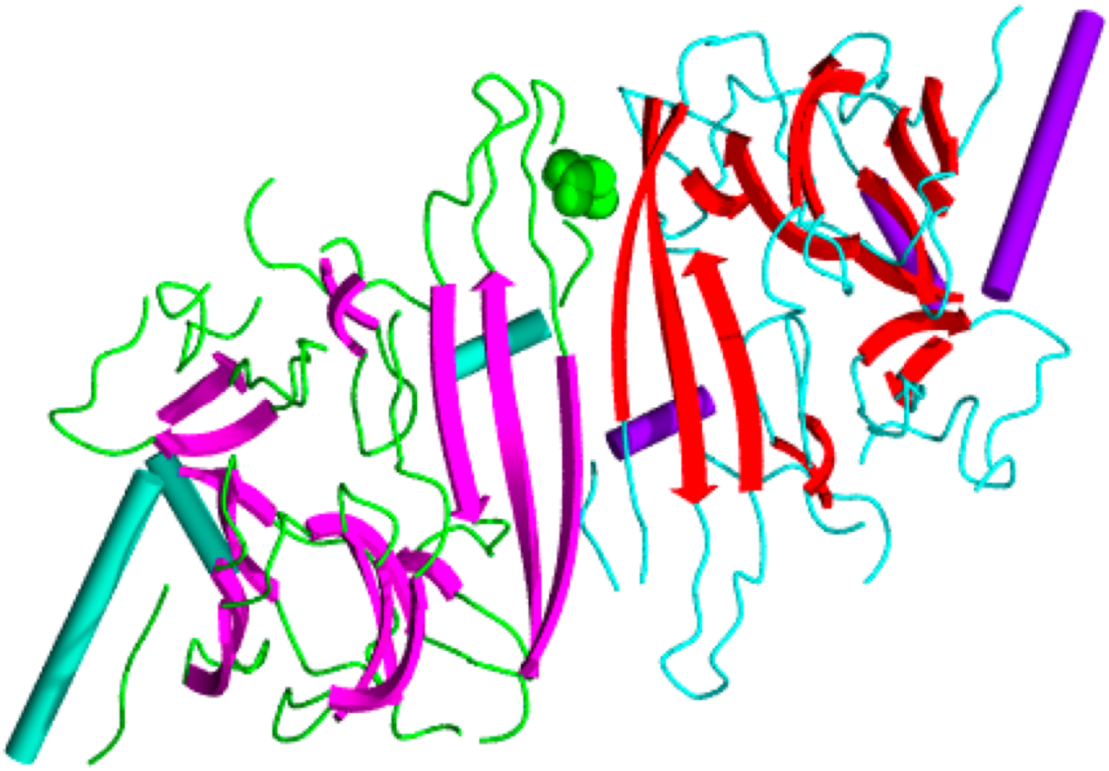

Figure S4B. Predicted structure of the ERN2 protein dimer, with the position of the difference between Spirit and Cedric indicated by green spheres.

4. Mining of metabolic pathways

Methods:

One hundred thirty-eight SAPs were found to be unique to tumor (Table S10). To be more precise, SAMTools called a SNP at that position based solely on reads from the tumor, while only one of the two alleles was observed in any of the reads from Cedric or Spirit. These SAPs were analyzed using the following pipeline. SAPs modifying the functions of proteins were determined by SIFT (12). The orthologs of genes with modifying mutations were traced to pathways in the *Monodelphis domestica* KEGG database (14-16). We converted 87 metabolic pathways of *Monodelphis domestica* into directed graphs with nodes representing reactants, products and proteins (i.e. gene products), and with edges representing the interactions among them. This kind of diagram captures the consecutive directional transformation of reactants into products (by the actions of proteins), which is characteristic of a pathway. It might be expected that the deletion of key nodes from the graph would cause its disconnection (just as the mutation of key genes results in metabolic pathway dissection). To predict the impact of gene mutations on tumor metabolism, the numbers of strongly connected components were compared before and after the deletion of nodes representing genes with modifying mutations.

monDom5	*hg19*	*gene*	*Tumor 1*	*Tumor 2*	*monDom*	*human*	*SIFT*
chr1:15464825	chr10:102053073	*PKD2L1*	H	Q	Q	Q	TOLERATED
chr1:15494467	chr10:101993071	*CWF19L1*	Q	L	Q	Q	TOLERATED
chr1:20121607	chr15:73541442	*NEO1*	S	T	S	A	TOLERATED
chr1:111400382	chr10:102775492	*PDZD7*	R	Q	R	Q	TOLERATED
chr1:172129473	chr14:23344678	*LRP10*	C	W	C	C	DELETERIOUS
chr1:183168644	chr15:52521333	*MYO5C*	T	I	T	V	DELETERIOUS
chr1:199736292	chr15:42600488	*GANC*	A	E	A	A	TOLERATED
chr1:203064470	chr15:45399113	*DUOX2*	V	L	A	A	TOLERATED
chr1:263260156	chr14:61924338	*PRKCH*	L	Q	L	L	DELETERIOUS
chr1:271594443	chr14:67854952	*PLEK2*	G	R	G	G	TOLERATED
chr1:292532223	chr5:135561034	*TRPC7*	R	H	R	R	TOLERATED
chr1:325555094	chr14:96769419	*ATG2B*	P	L	P	P	DELETERIOUS
chr1:343458529	chr5:150051947	*MYOZ3*	H	Q	H	H	TOLERATED
chr1:381139196	chr5:175110678	*HRH2*	T	M	T	T	DELETERIOUS
chr1:381828409	chr5:176008417	*CDHR2*	N	K	N	H	TOLERATED
chr1:382822619	chr5:176665375	*NSD1*	G	R	G	G	DELETERIOUS
chr1:390529386	chr20:42205016	*SGK2*	F	I	F	F	TOLERATED
chr1:443668824	chr16:53692369	*RPGRIP1L*	K	E	E	D	TOLERATED
chr1:450126859	chr20:45022193	*ELMO2*	G	S	G	G	TOLERATED
chr1:515946523	chr2:32712818	*BIRC6*	S	N	S	N	TOLERATED
chr1:561919028	chr2:69267186	*ANTXR1*	V	D	V	V	DELETERIOUS
chr1:668028539	chr16:78133744	*WWOX*	E	K	E	E	DELETERIOUS
chr1:696068893	chr16:88893155	*GALNS*	G	S	G	G	DELETERIOUS
chr1:717824437	chr2:85826309	*TMEM150A*	E	D	D	D	TOLERATED
chr1:719159520	chr2:84932862	*DNAH6*	A	S	A	A	TOLERATED
chr2:7305402	chr1:47755199	*STIL*	E	D	E	E	TOLERATED
chr2:104474500	chr1:201012421	*CACNA1S*	C	R	C	H	TOLERATED
chr2:104474518	chr1:201012439	*CACNA1S*	A	T	T	S	TOLERATED
chr2:160355913	chr1:231483648	*C1orf124*	R	H	R	S	TOLERATED
chr2:171538757	chr17:59946429	*INTS2*	R	K	R	R	DELETERIOUS
chr2:193623813	chr17:40826381	*PLEKHH3*	G	C	G	V	DELETERIOUS
chr2:198161479	chr17:37809911	*STARD3*	R	H	R	R	DELETERIOUS
chr2:286995005	chr6:42854103	*RPL7L1*	R	C	R	R	DELETERIOUS
chr2:287093745	chr6:42930055	*GNMT*	K	E	E	E	TOLERATED
chr2:287104485	chr6:42936201	*PEX6*	A	V	G	G	DELETERIOUS
chr2:376234620	chr6:110713993	*DDO*	S	F	S	P	TOLERATED
chr2:376295811	chr6:110777812	*SLC22A16*	A	V	A	A	TOLERATED
chr2:378353715	chr6:112394058	*TUBE1*	Q	H	Q	Q	DELETERIOUS
chr2:378453255	chr6:112471714	*LAMA4*	R	K	R	R	TOLERATED
chr2:429358963	chr6:150001210	*LATS1*	G	D	G	G	TOLERATED
chr2:440092005	chr6:158567901	*SERAC1*	L	F	L	V	TOLERATED
chr2:465133685	chr1:99380414	*LPPR5*	D	G	D	D	DELETERIOUS
chr2:466529820	chr1:100349752	*AGL*	L	I	I	I	TOLERATED
chr2:488798397	chr1:117503898	*PTGFRN*	N	K	K	K	TOLERATED
chr2:493655471	chr1:147096376	*BCL9*	A	T	P	P	TOLERATED
chr2:494438993	chr1:145441189	*TXNIP*	P	H	P	P	DELETERIOUS
chr2:506388037	chr17:26851751	*FOXN1*	S	G	S	S	DELETERIOUS
chr2:517241151	chr17:1656055	*SERPINF2*	M	I	M	M	TOLERATED
chr2:521608426	chr2:231944268	*PSMD1*	A	P	A	A	TOLERATED
chr2:523813440	chr21:47961690	*DIP2A*	P	T	P	P	TOLERATED
chr3:26660434	chr5:61643953	*KIF2A*	P	L	P	P	TOLERATED
chr3:63622338	chr18:43206954	*SLC14A2*	I	V	V	V	TOLERATED
chr3:63640420	chr18:43219827	*SLC14A2*	A	T	A	A	DELETERIOUS
chr3:101320166	chr5:6652006	*SRD5A1*	I	V	I	A	TOLERATED
chr3:162064630	chr8:75263599	*GDAP1*	R	H	R	R	DELETERIOUS
chr3:223010891	chr5:83239272	*EDIL3*	R	W	R	R	DELETERIOUS
chr3:240127122	chr5:36251526	*RANBP3L*	K	N	K	K	DELETERIOUS
chr3:243277288	chr5:33461302	*TARS*	I	T	T	T	TOLERATED
chr3:287631288	chr18:64211295	*CDH19*	L	I	V	V	TOLERATED
chr3:429210520	chr19:12880800	*HOOK2*	L	Q	L	Q	TOLERATED
chr3:431454301	chr19:10104064	*COL5A3*	D	E	D	D	TOLERATED
chr3:444936562	chr19:10132111	*RDH8*	E	Q	E	Q	TOLERATED
chr3:452498124	chr19:19162829	*ARMC6*	G	R	G	G	TOLERATED
chr3:475364342	chr12:125437112	*DHX37*	A	V	T	N	TOLERATED
chr3:488082422	chr12:124798747	*FAM101A*	E	K	E	E	TOLERATED
chr3:488626726	chr12:109042370	-	L	M	M	-	TOLERATED
chr3:488781243	chr12:108929161	*SART3*	M	T	M	M	DELETERIOUS

chr3:496270850	chr12:113403578	*OAS3*	R	*	R	R	DELETERIOUS
chr4:73395860	chr3:111831973	*C3orf52*	A	T	A	T	TOLERATED
chr4:75923846	chr3:113850134	*DRD3*	A	D	D	E	TOLERATED
chr4:110428436	chr2:128350428	*MYO7B*	K	E	K	H	DELETERIOUS
chr4:205713997	chr2:189906385	*COL5A2*	E	*	E	E	DELETERIOUS
chr4:216834521	chr2:196753147	*DNAH7*	C	Y	C	C	DELETERIOUS
chr4:223079623	chr11:124078631	-	V	G	V	-	DELETERIOUS
chr4:259554509	chr11:102818674	*MMP13*	V	I	V	V	TOLERATED
chr4:337299797	chr11:76506658	-	T	M	T	-	TOLERATED
chr4:337597086	chr11:76415330	-	A	V	V	-	TOLERATED
chr4:350194427	chr11:4825325	*OR52R1*	M	I	I	I	TOLERATED
chr4:351385972	chr11:5221019	*OR51V1*	R	Q	R	K	DELETERIOUS
chr4:352790000	chr11:5566483	*OR52H1*	E	K	E	L	NOT_SCORED
chr4:352802628	chr11:5573175	-	V	I	T	-	TOLERATED
chr4:353262978	chr11:5740852	-	L	M	M	-	TOLERATED
chr4:353263344	chr11:5741217	-	E	K	E	-	TOLERATED
chr4:353263518	chr11:5741391	-	I	V	I	-	TOLERATED
chr4:377715188	chr19:51830993	*IGLON5*	G	V	G	S	TOLERATED
chr4:382647171	chr19:39017654	*RYR1*	R	Q	R	R	TOLERATED
chr4:385795074	chr19:47861324	*DHX34*	R	H	R	R	TOLERATED
chr4:385870625	chr19:47764015	*CCDC9*	R	W	R	R	DELETERIOUS
chr4:394113814	chr1:6171937	*CHD5*	H	Y	H	H	DELETERIOUS
chr4:433410708	chr1:9324280	*H6PD*	W	R	R	R	TOLERATED
chr5:7961272	chr4:122741831	*CCNA2*	E	G	E	E	DELETERIOUS
chr5:15254925	chr8:17868138	*PCM1*	L	F	L	H	TOLERATED
chr5:29696473	chr4:94436442	*GRID2*	R	C	R	R	DELETERIOUS
chr5:42331418	chr4:107183321	*TBCK*	Q	L	Q	Q	TOLERATED
chr5:61572215	chr4:79396641	*FRAS1*	T	I	T	T	TOLERATED
chr5:63480416	chr4:111470867	-	S	Y	S	-	TOLERATED
chr5:66053747	chr4:113553113	*C4orf21*	I	T	I	I	DELETERIOUS
chr5:120759051	chr4:156717637	*GUCY1B3*	S	N	S	S	TOLERATED
chr5:122734474	chr4:155219317	*DCHS2*	Q	K	Q	E	TOLERATED
chr5:224024499	chr4:6611731	*MAN2B2*	M	I	I	M	TOLERATED
chr5:224027480	chr4:6612911	*MAN2B2*	F	L	L	L	TOLERATED
chr5:224361834	chr4:6864738	*KIAA0232*	A	V	A	A	DELETERIOUS
chr5:255803218	chr11:20119179	*NAV2*	L	V	L	L	TOLERATED
chr5:284961680	chr11:48157608	*PTPRJ*	S	N	D	D	TOLERATED
chr5:293430943	chr11:124135047	*OR8G5*	S	F	S	S	DELETERIOUS
chr5:298971794	chr11:60615417	*CCDC86*	D	E	D	D	DELETERIOUS
chr5:301696151	chr11:3039918	*CARS*	A	T	A	A	TOLERATED
chr6:97529456	chr9:71992638	*FAM189A2*	C	Y	C	C	DELETERIOUS
chr6:155931909	chr16:1568265	*IFT140*	V	D	V	V	DELETERIOUS
chr6:158647846	chr16:29998854	*TAOK2*	V	I	V	V	DELETERIOUS
chr6:198873235	chr1:228238460	*WNT3A*	T	P	T	S	TOLERATED
chr6:234421610	chr3:3143441	*IL5RA*	V	I	I	I	TOLERATED
chr6:252107685	chr3:127396104	*ABTB1*	D	E	D	D	DELETERIOUS
chr6:289457141	chr7:36375800	*KIAA0895*	V	I	V	V	DELETERIOUS
chr7:32334922	chrX:10104741	*WWC3*	R	*	R	R	DELETERIOUS
chr7:62131896	chr2:100915757	*LONRF2*	N	D	S	N	TOLERATED
chr7:174479509	chr2:219128116	*GPBAR1*	W	R	R	R	TOLERATED
chr7:174479589	chr2:219128036	*GPBAR1*	Q	R	Q	Q	TOLERATED
chr7:184227650	chr2:211476974	*CPS1*	E	D	E	E	TOLERATED
chr7:193462370	chr2:203826092	*ALS2CR8*	R	L	R	R	DELETERIOUS
chr7:253818548	chr3:193376686	*OPA1*	Q	H	Q	Q	DELETERIOUS
chr7:254003289	chr3:193128777	*ATP13A4*	R	Q	R	R	TOLERATED
chr8:10587931	chr22:45790601	*SMC1B*	R	C	R	R	DELETERIOUS
chr8:11687886	chr22:46782462	*CELSR1*	S	N	S	S	TOLERATED
chr8:28253121	chr12:49953531	*MCRS1*	V	G	V	V	DELETERIOUS
chr8:50792459	chr12:81102314	*MYF6*	D	H	D	D	TOLERATED
chr8:64548504	chr22:39824138	*TAB1*	G	S	G	G	DELETERIOUS
chr8:91118185	chr22:36897350	*FOXRED2*	A	P	A	A	DELETERIOUS
chr8:108711718	chr12:7303196	*CLSTN3*	R	Q	R	R	TOLERATED
chr8:108720900	chr12:7310644	*CLSTN3*	R	C	R	R	DELETERIOUS
chr8:116310091	chr12:2994655	-	C	R	H	-	TOLERATED
chr8:119670529	chr12:2602519	*CACNA1C*	T	A	T	T	DELETERIOUS
chr8:189596823	chr7:129357088	*NRF1*	G	S	G	G	DELETERIOUS
chrX:3278457	chrX:153669478	*GDI1*	N	K	N	N	TOLERATED
chrX:5285219	chr5:178339710	-	V	L	T	-	TOLERATED
chrX:5868111	chrX:109441761	*AMMECR1*	N	D	N	N	DELETERIOUS

chrX:51183433	-	-	I	M	I	-	DELETERIOUS
chrX:71809857	chr2:131103657	*IMP4*	V	A	A	A	TOLERATED

Table S10. Putative amino-acid variants observed only in the tumor. The columns give the corresponding location in the *Monodolphis* genome assembly called "monDom5", the corresponding location in the human genome assembly called "hg19" (if it could be determined), the name of the putatively orthologous human gene, the predicted variant amino acids and their *Monodelphis* and human orthologs, and the prediction made by the SIFT program whether the polymorphism would be deleterious or tolerated if it occurred in the *Monodolphis* protein.

Results:

Of the initial 138 SAPs, 54 were found to be modifying mutations. Seven of these 54 were traced to 24 pathways, in which we observed one graph disconnection. This disconnection was produced in the glycosaminoglycan degradation metabolic pathway by a modifying mutation in the gene GALNS (Figure S5). Tables S11 and S12 show the genes and pathways with the highest number of SAPs and modifying mutations. A suitable structural model was found for three genes potentially associated to cancer (Table S13). The most interesting findings about these three genes are explained next.

PRKCH (PKCH). Protein kinase Cη (*prkcη*) is an enzyme with multifunctional catalytic activity involved in the transduction of signals for cell growth and differentiation (17, 18). This enzyme is expressed in the granular layer of the epidermis and is activated by calcium and phorbol esters or DAG. Interestingly, its expression has been associated with tumor suppression in skin cancer (17, 18).

In protein kinases the catalysis is driven by a catalytic motif in concert with one activation loop and one α helix (i.e. "αC helix") (19). The activation loop binds the substrate, and the αC helix docks cofactors. In the phosphorylated conformation of *prkcη*, the residues Thr-500, Arg-465 and Lys-489 position the catalytic residue Asp-466 for substrate catalysis and align Glu-490 and Glu-390 to form hydrogen bonding with Arg-392 and Lys-371 respectively. In parallel Arg-392 forms a salt bridge with Glu-385 (19). This general conformation aligns and stabilizes substrate for catalysis (Figure S6).

The mutation L394Q occurring in tumor is located in αC helix in proximity of Glu-390 and Arg-392. The change from a nonpolar residue to a polar one would destabilize the αC helix structure and avoid the formation of hydrogen and salt bridges of residues Glu-390 and Arg-392. As a result, the alignment of substrate will be distorted as will the catalysis reaction. With the current data, it is difficult to predict the energetic changes in catalysis; however we speculate that the deficient alignment of the substrate will result in a deficient catalysis and a more costly (i.e. endergonic) reaction.

GALNS. Galactosamine (N-acetyl)-6-sulfate sulfatase enzyme (*galns*) hydrolyzes sulfate ester bonds of GalNAc-6-S and Gal-6-S at the terminal of chrondroitin-6-sulfate and keratan sulfate respectively (20). Deficiency of this enzyme results in Morquio's syndrome (21).

The closest structural model to *galns* is the model for the human lysosomal arylsulfatase A (*asa*) (Table S13). The catalytic site of this protein comprises the residues Asp-29, Gly-69, Asp-281, Lys-302 and Lys-123. In addition to these residues, high-mannose-type oligosaccharide side chains are attached to residues Asn-158, Asn-184, and Asn-350. These residues act as lysosomal targeting signals that mediate *galns* vesicular transportation from Golgi

compartments to the lysosomes (22, 23). It has been reported that mutations in the residue Asn-350 ends in a shorter aberrant form of the protein with reduced activity (24).
GALNS genes of Cedric, Spirit and tumor code for Asp instead of Asn in position 350.. Despite this, we speculate that this residue might play a similar role in *galns* and in *asa*, as suggested by its structural conservation and the chemical similarity of Asn and Asp. In contrast, the residue Gly-355 is conserved in different sulfatases (and species) but not in tumor (Figure S7a). This mutation (i.e. G355S) results in a residue with a greater surface area and a lower hydrophobicity. It is not possible to predict with certainty the energetic effects of the amino-acid substitution, yet we speculate that the neighbour residues would rearrange as a result of the displacement produced by the introduction of the NH3 group of serine (this includes Ile-353). Thus the mutation G355S introduced in tumor might result in the displacement of Asp-350 (by means of Ile-353) and a consequent deficient glycosylation, phosphorilation and transport of *galns*.

CCNA-like. Cyclin A (*ccna*) is involved in cell cycle regulation. This protein is made up of 12 alpha-helices, five of which have conserved residues that interact with *cdk2*. On the opposite side of this interface, a second cluster of conserved residues (denominated "Cyclin A CLS") is located (25, 26). Following the description of Brown et al. (27) this region has in its core the residues Trp-217 and Gln-254. Surrounding Trp-217 four hydrophobic residues (Val-221, Leu-218, Leu-214, and Ile-213), and other amino acids from helix 1 (Val-215, Asp-216, Val-219 and Glu-223) can be found (Figure S8). In the vicinity of Gln-254, a negatively charged cluster is created by residues Glu-220, Glu-223 and Glu-224. Additionally, the residues 210-214 form a highly conserved motif (MRAIL), which is expected to influence the interactions of CLS. A recent experiment proposed that CLS functionality depends on four solvent-accessible residues including Glu-220 (26).

In the tumor, the change from a polar negative residue (Glu) to a non-polar neutral residue (Gly) in position 220 impacts the function of *ccnA-like*. This mutation modifies the patch of negative residues and recreates the results of the experiment of Pascreau et al. (26). In this way, the mutation E220G will deprive the *ccnA-like* protein in tumor cells from a precise control of DNA replication.

Mutations potentially associated to cancer pathology. The methodology we used pointed to three genes that might be related to cancer pathology. Further studies should determine the role of these mutations in cancer transmission and proliferation. In addition to these three, other mutations found in Devil's tumor cells might be related to cancer. By comparing multiple polygenomic tumors a recent study associated focal deletions and amplifications of gene CACNA1 to cancer progression (28). Similarly, deleterious mutations in the gene NRF1 has been associated to the development of hepatic cancer in mice (29). The hypothetical effect of the mutations in the three genes related to cancer pathology is explained next.

Hypothesis 1. The activity of protein kinase Cη (prkcη) is deficient or null in tumor cells. This results in a deficient control of cell multiplication that originates and/or propagates cancer. PRKCH is expressed exclusively in the granular layer of the epidermis. Its product *(prkcη)* is an upstream activator of keratinocyte differentiation that also induces cell cycle arrest in association with *cdk2*, *p21* and cyclin E. In this way, functional *prkcη* should induce keratinocytes differentiation and control their multiplication in skin (i.e. epidermis) (18). Previous experiments suggest that *prkcη* suppresses tumor progression, controls induced epidermal hyperplasia, promotes wound healing and plays a role in the maintenance of epithelial architecture (17). We hypothesize that the mutated *prkcη* enzyme (with deficient catalysis) in tumor cells is unable to control tumor progression, keratinocyte differentiation, and

cellular growth. Thus *prkcη* might contribute to cancer formation and progression in the Tasmanian Devil.

Hypothesis 2. The degradation of keratan sulfate and chondroitin sulfate is null or inefficient in tumor cells. Its accumulation results in an aberrant regulation of CD44 and other growth factors that might end in cancer or a more aggressive form of cancer. In the tumor, the glycosaminoglycan degradation pathway was found to be dissected by one modifying mutation in the gene GALNS that decreases the activity of its product (*galns*). This dissection would result in an aberrant accumulation of keratan sulfate and chondroitin sulfate and their further abnormal incorporation into proteoglycans. Lumican is a small leucine-rich proteoglycan present in multiple tissues and forms of cancer (30). It has been reported that forms of human metastatic melanoma cell lines secrete lumican in a proteoglycan form mostly with keratan sulfate chains (31). In the same way, the presence of keratan sulfate has being correlated with the malignancy of astrocytic tumors (32). Other reports point out that keratan sulfate plays an important regulatory role of CD44, a factor involved in multiple biological processes including cell growth and tumor progression (33, 34). Just as for keratan sulfate, chondroitin sulfate is degraded by *galns*. Chondroitin sulfate is involved in morphogenesis, wound healing and growth factor recruitment. It has been reported that over-sulfated chondroitin chains strongly interact with growth factors involved in tumor growth and progression (i.e. midkine, PTN, FGF-16, FGF-18) (35).

We hypothesize that accumulation of keratan sulfate and chondroitin sulfate would result in their higher incorporation into proteoglycans, and a concomitant aberrant regulation of CD44 and other growth factors. They might result in aberrant cell growth and reduce the ability of the immune system to recognize tumor cells.

Hypothesis 3. Cyclin A-like has a homologous role to cyclin A. The mutated cyclin A-like competes with wild cyclin A and cyclin E causing a misallocation of DNA replication factors. In the tumor, DNA duplication would not be efficiently inhibited, resulting in uncontrolled and/or aberrant cell growth. In Eukaryotes, the cell cycle is finely controlled by cyclin-dependent kinases (CDKs) and cyclin subunits. DNA replication is initiated by the increased activity of *cdk2* during G1-S transition (25-27). DNA replication is arrested without *cdk*, while retarding *cdk2* activity results on a longer S phase and DNA multiplication (i.e. centrosome multiplication) (25, 26). *Cdk2* is mainly activated by cyclin E but can be activated by cyclin A as well (25, 26). Additionally cyclin A modulates *cdk2* specificity for p107, the transcription factor E2F and the replication protein-A (27).

Cyclin A CLS interacts with the DNA replication factors p27, *mcm5* and *orc1* (25). *Mcm5* inhibits centrosome amplification in S-phase-arrested CHO cells. *Orc1* is an essential component in DNA replication, while p27 removes cyclin A but not cyclin E from centrosomes. Expression of the cyclin A CLS displaces both endogenous cyclin A and E from centrosomes and inhibits DNA replication. In addition, by interacting with *mcm5* and *orc1*, cyclin A CLS inhibits the centrosome reduplication throughout S phase and G2 (25).

It has been proposed that an aberrant centrosome number is related to chromosome mis-segregation, genomic instability and tumor development (25). We hypothesize that the mutated cyclin A-like of tumor cells might compete against cyclin A and cyclin E for the available binding sites in centrosomes. This competition will result in a biased or insufficient positioning of p27, *mcm5* and *orc1*in centrosomes during S-phase resulting in a higher-than-expected DNA replication and posterior cell growing aberration.

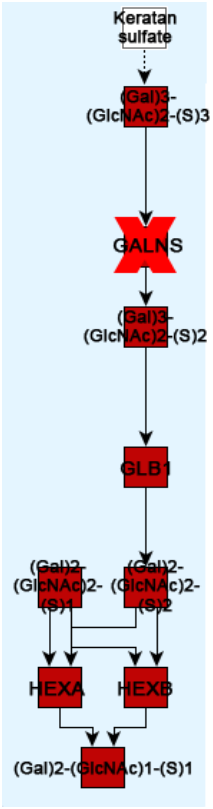

Figure S5. Keratan sulfate degradation graph. This is a part of the Glycosaminoglycan degradation pathway (KEGG mdo00531). The modifying mutations in the tumor gene *GALNS* (marked with X) were found to disconnect the pathway. This graph is expected to be connected in Cedric and Spirit.

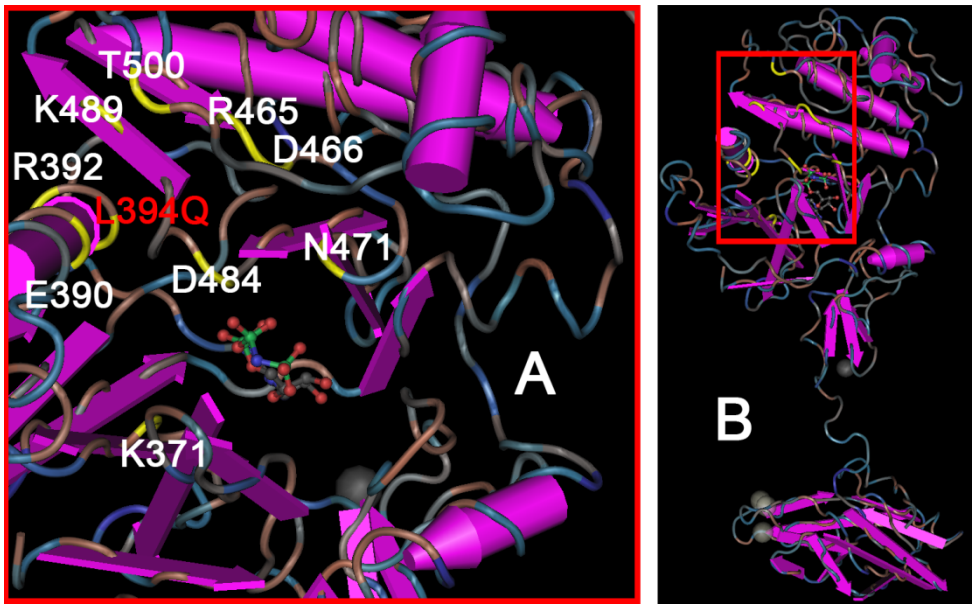

Figure S6. Mutation in protein kinase Cη (*pkcη*) occurring in tumor. Hydrophobic residues are plotted in blue. The mutation in tumor (red) is located in the αC helix, which works as a docking site for cofactors and plays an important role in the alignment of the substrate during catalysis (19). *Prkcη* suppresses tumor progression, controls induced epidermal hyperplasia, promotes wound healing and plays a role in the maintenance of epithelial architecture (17, 18).

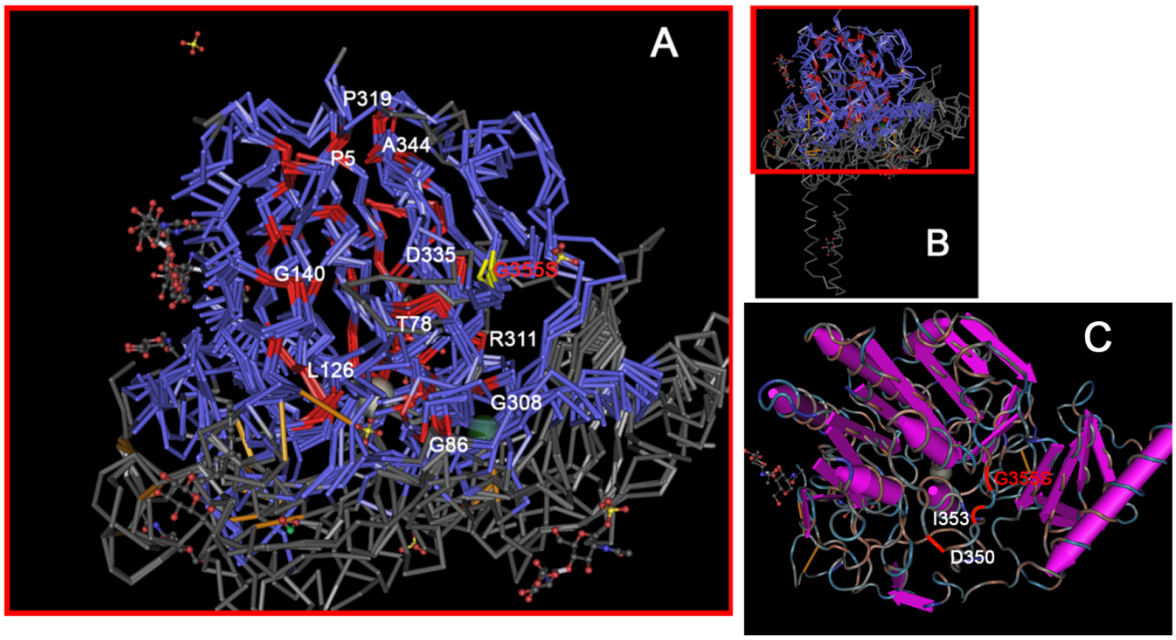

Figure S7. Mutation in galactosamine (N-acetyl)-6- sulfate sulfatese (*galns*) occurring in tumor. (A) Conserved residues in proteins with structural similarity to GALNS. Residues found to be identical in all the proteins (red) include the external residue mutated in tumor (G355S in yellow). (B) Proteins similar to *galns* are broadly distributed in life domains and tissues. (C) Residue Asp-350 is expected to be glycosylated and phosphorilated in normal cells. The mutation Ser-355), might affect the position of Asp350in tumor..

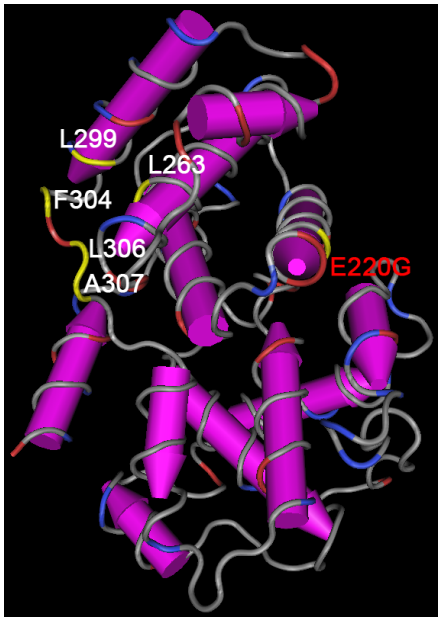

Figure S8. Mutation in cycline A-like (*ccnA-like*) protein occurring in tumor. Hydrophobic residues are plotted in blue. The mutations in tumor (red) is located in the opposite side of the cyclin A-CDK interface (yellow) in the CLS region. CLS regionmediates the contact with other proteins (i.e. p27, MCM5, Orc1) and plays an important role in prolonging the inhibition of centrosome reduplication throughout the cell cycle (25, 26).

Ortholog in *Monodelphis domestica*	Gene	Gene annotation	Mutations	Modifying Mutations
ENSMODT00000003592	MAN2B2	Mannosidase Alpha Class 2B	2	0
ENSMODT00000019680	GPBAR1	G protein-coupled bile acid receptor 1	2	0
ENSMODT00000037274	CLSTN3	Calsyntenin 3	2	1
ENSMODT00000037496	CACNA1S	Calcium channel, voltage dependent	2	0

Table S11. Annotated genes in tumor with the highest number of SAPs. Genes with SAPs in tumor were traced to their orthologs in *Monodelphis domestica*. The *Monodelphis domestica* orthologs were annotated with ENSEMBL codes. SAPs modifying the functions of proteins were determined by SIFT.

Pathway	Accumulated number of mutations	Genes with mutations
MAPK signaling pathway	2	CACNA1C, TAB1
Vascular smooth muscle contraction	2	PRKCH, CACNA1C
Cell cycle	2	CCNA-like, SMC1B

Table S12. Pathways in the tumor with the highest number of modifying SAPs. The orthologs of genes with modifying mutations were traced to pathways in the *Monodelphis domestica* KEGG database.

Orthologous in *Monodelphis domestica*	Gene name	PDB model	Bit score	Coverage percentage	E-val	Mutation position in model
ENSMODT00000004992	GALNS	1N2K_A	263	81%	1.00E-70	G355S
ENSMODT00000009166	SMC1B	2WD5_A	291	17%	9.00E-79	-
ENSMODT00000011053	PRKCH	3PFQ_A	565	88%	1.00E-112	L394Q
ENSMODT00000011972	TAB1	2J4O_A	734	79%	0	-
ENSMODT00000018631	NRF1	1GOJ_A	32	20%	0.74	-
ENSMODT00000023397	CACNA1C	3OXQ_E	159	4%	7.00E-39	-
ENSMODT00000024041	CCNA-like	1VIN	499	62%	7.00E-142	E220G

Table S13. Models used to predict structural changes in tumor. All the models belong to the PDB database. Models with low similarity were excluded from further analysis. Mutations were traced into models by BLAST. The mutation in gene ENSMODT00000011972 was not covered by the model.

5. Samples for mitochondrial sequencing and sequence generation.
The seven modern animals selected for full mitochondrial sequencing were selected based on geographical location and availability of DNA from hair shafts (Figure 2A, identified in red). They included, along with reference animals Spirit (Freycinet Peninsula) and Cedric (Woolnorth), animals captured from Mount William (TD200), Epping Forest (TD197), Mole Creek (Copenhagen zoo devil "Montague"), Narawntapu Wildlife Park (mainland captive bred animal "TDWilmur"), and Granville Harbour (TD184).

All sequencing on hair extracted DNA was carried out using the Roche GS FLX 454-sequencer with FLX chemistry (read length up to 250 bp). For mitochondrial targeted-resequencing PCR amplicons were analyzed using BigDye Terminator v3.1 (Applied Biosystems, Foster City, CA, USA) chemistry on an ABI 3730 genetic analyzer. Nine of the 13 mitochondrial genomes were sequenced using MID-tagged samples, which were pooled for simultaneous sequencing in a 2-chamber region gasket. Each MID was recognized by the Roche GS FLX analysis software as a unique sample. Between 6- and 175-fold coverage of individual Mitochondrial genomes were generated and sequences assembled as described for the Tasmanian tiger (10).

Genotyping of the 17 identified informative mitochondrial SNPs were performed by pre-sequencing PCR-based denaturing gradient gel electrophoresis (DGGE) and/or Sanger sequencing. DGGE was performed using the Ingeny phorU-2 DGGE system (Ingeny International, Goes, The Netherlands; www.ingeny.com) and gel conditions as previously described for broad range DGGE (36). Primer sequences with GC-clamp conditions are available on request (V.M.H.). All DGGE banding patterns were confirmed via Sanger sequencing. Amplified products were purified and fluorescently labeled using BigDye Terminator v3.1 (Applied Biosystems, Foster City, CA, USA) chemistry according to manufacturer's guidelines and sequenced on an ABI 3730 genetic analyzer (Applied Biosystems; www.appliedbiosystems.com) using standard procedures.

6. Mitochondrial diversity.

To get a sense of the relative sparsity of mitochondrial differences in *Sarcophilus*, we compared it with the analogous figure (Figure S9) for 20 randomly chosen mitochondria from Europeans, obtained at http://www.genpat.uu.se/mtDB/. For mitochondrial genomes of humans and other species (in Figure 1C in the main paper), self-alignments were used to locate hypervariable regions, which were excluded from the analysis.

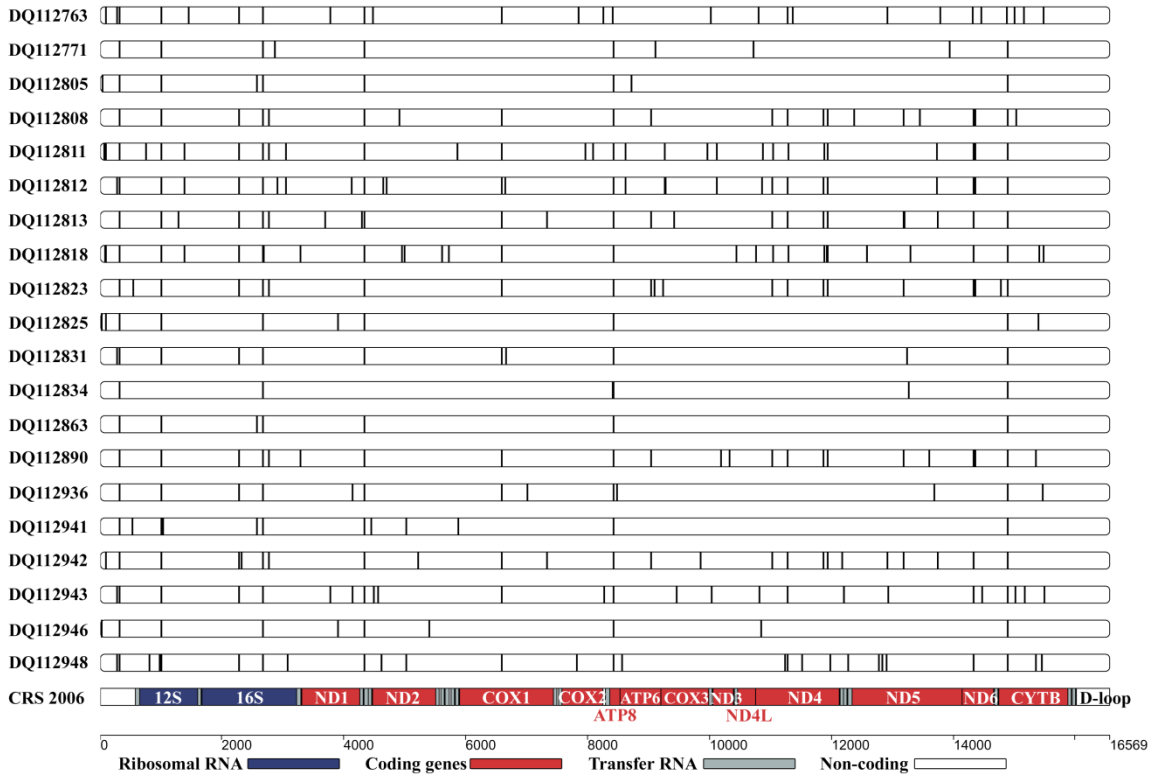

Figure S9. Positions of differences between the Cambridge reference human mitochondrial sequence and each of the 20 randomly selected European mitochondrial genomes.

Data for Figure 1C. The average numbers of nucleotide substitutions between two mitochondrial genomes are shown in Figure 1C of the main paper. For Europeans, we averaged over the 20 sequences in Figure S9. Data for the Bushmen comes from ref. (11). For the remaining columns of Figure 1C, we used the following GenBank entries: Gorilla: GORMTC (37) and X93347 (38); Wolf: DQ480503-DQ480508 (39) and AM711902 (40); Panda: EF212882 (41) and AM711896 (41); Whales: (*Balaenoptera brydei*; Bryde's whale) AP006469 (42) and AB201259 (36), (*Eschrichtius robustus*; grey whale) AP006471 (42) and AJ554053 (36), (*Balaenoptera acutorostrata*; minke whale) AP006468, AJ554054 (43), (*Balaenoptera omurai*; Omura's baleen whale) AB201256, AB201257 (36), (*Caperea marginata*; pigmy right whale) AP006475 (42), and AJ554052 (43); Woolly mammoth, clade I: EU153444-EU153449, EU153452, EU153454 -EU153458 (44), EU155210 (45), DQ188829 (46), DQ316067 (47); Woolly mammoth, clade II: EU153450, EU153451, EU153453 (44) Steller sea lion: AJ428578 (48) and AB300601-AB300608 (Yanagimoto,T. and Moriya,A., unpublished); Tasmanian devil: data from this paper; Tasmanian tiger: FJ51578, FJ515781 (10); American bison (49).

Table S14. Analysis of 17 informative mitochondrial SNPs in 175 geographically dispersed modern animals and 6 museum specimens.

	Informative SNPs (nt position)																	Frequency		
Haplo-group[1]	734	1091	5391	5462	6891	8124	9773	10235	12158	12510	12719	12721	14088	14125	14411	14645	16294			
	Modern animals (n=175)																	Reference Concordance[2]	East n=98	West n=77
A	**G**	**G**	**T**	**G**	**G**	**A**	**T**	**T**	**T**	**C**	**C**	**C**	**C**	**C**	**T**	**T**	**G**	100%	50%	1%
B	G	A	T	G	G	A	T	T	T	T	C	C	C	T	T	T	G	82%	7%	8%
C	G	A	T	G	G	A	T	T	T	T	C	T	T	T	T	T	G	71%	36%	5%
D	G	A	T	A	G	G	T	T	T	T	T	C	T	T	T	C	G	53%	7%	0%
E	A	A	C	G	A	A	C	C	C	T	C	T	C	T	C	T	A	29%	0%	86%
	Museum specimens (n=6)																	Concordance[3]	Number of specimens	Collection Dates
A	G	G	T	G	G	A	T	T	T	C	C	C	C	C	T	T	G	100% A	2	1991, 1994
hA	G	G	T	G	G	A	T	T	T	T	C	C	C	C	T	T	G	94% A	1	1908
hF	G	A	T	G	G	A	T	T	T	T	T	C	T	T	T	T	A	new	1	1870-1910
E	A	A	C	G	A	A	C	C	C	T	C	T	C	T	C	T	A	100% E	2	1930, 1931

[1] Modern haplogroups A to E; reference haplogroup in bold; h, historic haplogroup
[2] Percentage haplogroup concordance with Reference east sequence (Spirit), haplogroup A.
[3] Percentage haplogroup concordance with closest modern haplogroup

7. Genotyping.

Sampling was based on previously published criteria, defined as the core distributional range of devils (Figure S10), predicted by measures of primary habitat including rainfall, terrain and major prey (50). Animal selection was random, based on maximizing collection area (i.e. distance between collection animals), with the aim of minimizing possible kinship (in communication M.J.). We included 87 devils (including Spirit and Cedric) from 12 major study sites (averaging 8 animals per site) of varying capture sizes for comparison of mitochondrial and nuclear diversity. An additional 89 animals (including captive animals, see below) and 7 capture sites were included for exhaustive analysis of current mitochondrial haplotype dispersal (n=175). From the northeastern corner of Tasmania, the origin of disease outbreak, moving south along the coast, northeast across central Tasmania and up to the northwestern corner where the disease prevalence is at its lowest, the following capture sites and geographical collection ranges were included; Mount William (25 km^2), St Helen's to St Mary's coastal region (200 km^2), Mole Creek included a single collection of 4 disease-free animals from Trowunna Wildlife Park and sent to Prince Frederick and Tasmanian born Princess Mary of Denmark as a wedding gift from Tasmania in 2006 (currently located at the Copenhagen Zoo), Freycinet Peninsula (160 km^2), Little Swanport (extended 25 km^2 inland), Pawleena (a single disease-free animal), Forestier Peninsula (150 km^2), moving inland Epping Forest (a single disease-free animal), Fentonbury (25 km^2), Bronte National Park (25 km^2), Lake Rowallan region at the current disease front (25 km^2), the largest collection region stretching from Strahan to Corinna along the west coast and inland to Luina (500 km^2) is divided into the Strahan and Luina collections, Temma to Marrawah coastal region (400 km^2), Dip Falls (400 km^2), Milkshake Hills (25 km^2), and Woolnorth (400 km^2). Capture ranges (25 to 500 km^2) were

adopted to reduce the effects of sampling bias, based on the predicted 10-20 km^2 life-time home range estimate (51). Trapping and sampling (by an ear tissue biopsy) was conducted either in 2005 for the majority of the eastern devils or in 2007 for the majority of the western devils, the latter as part of the selection for insurance population devils (www.tassiedevil.com.au). Tissue samples were stored in 70% ethanol at -20^oC and DNA extracted for the purpose of this study using the Qiagen DNA mini kit (Qiagen Inc., Hilden, Germany).

Captive Sampling. The 7 captive devils forming part of the insurance breeding program were sampled from the Australian Reptile Park in Gosford, New South Wales, on the mainland of Australia. Three of the recently acquired devils were captured in Woolnorth as part of the insurance intake and are therefore recent acquisitions to the park. These three animals were classified both geographically and via mitochondrial analysis as Woolnorth devils. The remaining devils are offspring of captured either pre-DFTD or insurance population devils mainly from the Narawntapu National Park region. Pedigree information and hair samples were obtained.

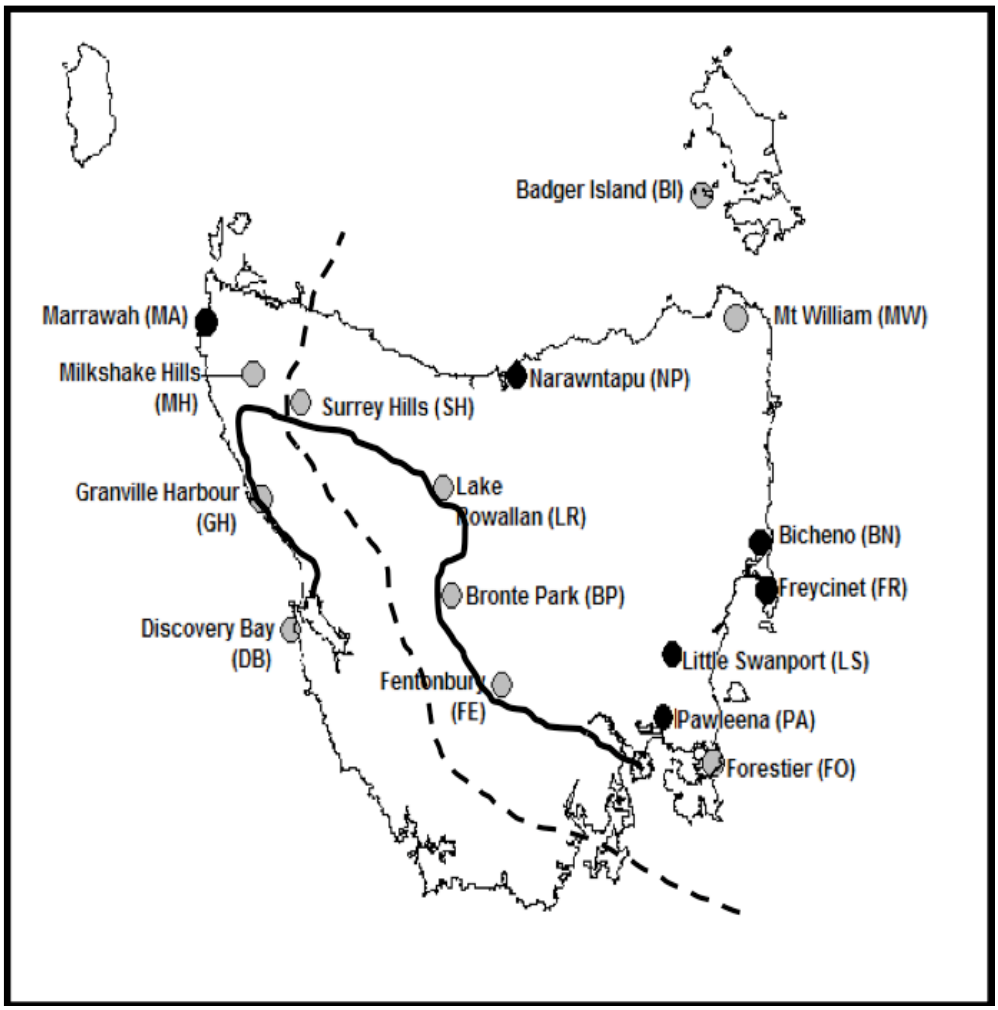

Figure S10. Map of Tasmania indicating the distribution of Tasmanian devils north of the solid line, as per estimations based on measures of primary habitat including rainfall, terrain and major prey. East of the dotted line indicates the disease front as per 2008 estimates.

8. Population structure.

Table S15 reports pairwise F_{ST} values between Tasmanian devils from 12 geographically defined populations. Note that the numbers cannot be directly compared with values obtained in other studies, because we selected SNPs based on their ability to distinguish among individuals,

Table S15. Pairwise F_{ST} between 12 geographically defined Tasmanian devil populations across Tasmania (n=87).

	Mt Haplogroup			921 Nuclear SNPs										
	H	**M**	**L**	**WN**	**MH**	**DF**	**TM**	**LN**	**SH**	**LR**	**BP**	**FB**	**FS**	**FC**	**MW**
Woolnorth (WN)	E	-	-	0.000	**-0.008**	0.018	0.012	0.084	0.119	0.175	0.201	0.217	0.308	0.264	0.289
Milkshake Hills (MH)	E	-	-	**-0.008**	0.000	**-0.054**	0.004	0.024	0.065	0.097	0.126	0.149	0.226	0.171	0.193
Dip Falls (DF)	E	-	C	0.018	**-0.054**	0.000	0.014	0.036	0.083	0.129	0.155	0.175	0.253	0.209	0.228
Temma (TM)	E	-	-	0.012	0.004	0.014	0.000	0.070	0.131	0.161	0.199	0.214	0.293	0.253	0.275
Luina (LN)	E	-	-	0.084	0.024	0.036	0.070	0.000	0.068	0.099	0.110	0.137	0.205	0.183	0.202
Strahan (SH)	B	-	-	0.119	0.065	0.083	0.131	0.068	0.000	0.076	0.087	0.118	0.198	0.177	0.198
Lake Rowallan (LR)	B	C	A	0.175	0.097	0.129	0.161	0.099	0.076	0.000	0.015	0.031	0.103	0.083	0.107
Bronte Park (BP)	C	-	-	0.201	0.126	0.155	0.199	0.110	0.087	0.015	0.000	**-0.000**	0.072	0.052	0.082
Fentonbury (FB)	C	A	B	0.217	0.149	0.175	0.214	0.137	0.118	0.031	**-0.000**	0.000	0.077	0.061	0.083
Forestier (FS)	A	-	-	0.308	0.226	0.253	0.293	0.205	0.198	0.103	0.072	0.077	0.000	0.051	0.051
Freycinet (FC)	A	-	C	0.264	0.171	0.209	0.253	0.183	0.177	0.083	0.052	0.061	0.051	0.000	0.015
Mount William (MW)	A	-	C	0.289	0.193	0.228	0.275	0.202	0.198	0.107	0.082	0.083	0.051	0.015	0.000

Number of animals per site: WN=11, MH=2, DF=8, TM=9, LN=4, SH=4, LR=8, BP=8, FB=8, FS=9, FC=8, MW=8. Shaded areas depict genetically similar populations with inbreeding (negative values) in some cases.

Notes on Figure 3C. Current methods of insurance animal selection are based on a random sampling that does not take population structure into consideration. Using that model, animals are ideally selected at equal contributions across the current habitat. As a result of rapid disease spread this may be skewed towards the non-disease region.

Analysis of 87 animals matched for mitochondrial and genome-wide genotyping data were used to define geographical 'collection zones' as alternative models for selection of insurance animals. Based on the mitochondrial model, which identified four major haplogroups, we would advise random sampling of insurance animals with equal percentage contributions from each of these zones. Our genome-wide genotyping approach allowed for more in-depth characterization of population structure and the identification of seven zones based on population inbreeding and minimal genetic diversity as defined by close-to-zero F_{ST} values (Table S15). Using this model we suggest that selecting equal percentages from each collection zone will allow for

maintaining maximal genetic diversity within the insurance population. Based on this model it is paramount that sampling not be restricted to the limited region of disease-free animals, but that active sampling within the diseased region be maintained and provide at least five-sevenths of the animal selected.

9. Conclusion. In this study we predict the impact of the limited genome-wide diversity on the survival of the Tasmanian devil whose survival from a contagious transmissible cancer disease will be placed in the hands of a restricted founder population that manages to naturally or within captivity survive DFTD. We demonstrate that complete saturation of mitochondrial DNA sequencing, although informative, reveals limited population structure compared with randomly selected dimorphic nuclear markers. Of major concern is the low genetic diversity and observed inbreeding within the only remaining viable disease-free Tasmanian devil sub-population in the northwestern region. Suggestions to isolate the northwestern population in the wild (via fencing and/or restricting insurance sourcing from this region) will result in a genetically limited inbred future population with reduced fitness. Inbreeding depression results in more recessive deleterious traits and increased risk of generating unviable offspring. The sensitivity of our approach resulted in the identification of two unique sub-populations south of this core region and ahead of the current disease front. These genetically independent sub-populations, along with sub-populations within the disease zone should be immediately sourced as selected founders to boost current insurance populations thus enhancing outbreeding. The rapid spread of the disease for over a decade has resulted in isolated sub-populations within the infected zone as Tasmanian devil numbers dwindle, further enhancing inbreeding depression within these groups. In conclusion, we highlight in this study that limiting the future Tasmanian devil population to northeastern derived founder animals, based on the assumption that the northwestern Tasmanian devil population may confer a resistance to the eastern-derived cancerous disease, is a major concern for the viability of this species from future natural challenges.

10. References

1. Myers EW, et al. (2000) A whole-genome assembly of Drosophila. *Science* 287(5461):2196-2204.
2. Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24(24):2818-2824.
3. Denisov G, et al. (2008) Consensus generation and variant detection by Celera Assembler. *Bioinformatics* 24(8):1035-1040.
4. Levy S, et al. (2007) The diploid genome sequence of an individual human. *Plos Biol.* 5(10):2113-2144.
5. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
6. Murphy WJ, Eizirikb E (2009) Placental mammals (Eutheria). *The Timetree of Life*, eds Hedges SB & Kumar S (Oxford University Press, Oxford), pp 471-474.
7. Springer MS, Krajewski C, Meredith RW (2009) Marsupials (Metatheria). *The Timetree of Life*, eds Hedges SB & Kumar S (Oxford University Press, Oxford), pp 471-474.
8. Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51(5):689-702.
9. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Phylogenet. Evol.* 15(12):1647-1657.

10. Miller W, et al. (2009) The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Res.* 19(2):213-220.
11. Schuster SC, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463(7283):943-947.
12. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31(13):3812-3814.
13. Miller W, et al. (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456(7220):387-U351.
14. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. (Translated from eng) *Nucleic Acids Res.* 28(1):27-30 (in eng).
15. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. (Translated from eng) *Nucleic Acids Res.* 38(Database issue):D355-360 (in eng).
16. Kanehisa M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. (Translated from eng) *Nucleic Acids Res.* 34(Database issue):D354-357 (in eng).
17. Chida K, et al. (2003) Disruption of Protein Kinase C eta results in impairment of wound healing and enhancement of tumor formation in mouse skin carcinogenesis. *Cancer Res.* 63(10):2404-2408.
18. Kazanietz MG, Denning MF (2010) PKC isozymes and skin cancer. *Protein Kinase C in Cancer Signaling and Therapy,* Current Cancer Research, ed El-Deiry W (Humana Press), pp 323-345.
19. Grodsky N, et al. (2006) Structure of the catalytic domain of human protein kinase C beta II complexed with a bisindolylmaleimide inhibitor. *Biochemistry* 45(47):13970-13981.
20. Masue M, Sukegawa K, Orii T, Hashimoto T (1991) N-Acetylgalactosamine-6-Sulfate sulfatase in human placenta: Purification and Characteristics. *J. Biochem.* 110(6):965-970.
21. Di Ferrante N, Ginsberg LC, Donnelly PV, Di Ferrante DT, Caskey CT (1978) Deficiencies of Glucosamine-6-Sulfate or Galactosamine-6-Sulfate sulfatases are responsible for different mucopolysaccharidoses. *Science* 199(4324):79-81.
22. Lukatela G, et al. (1998) Crystal structure of human Arylsulfatase A: The aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. *Biochemistry* 37(11):3654-3664.
23. Schierau A, et al. (1999) Interaction of Arylsulfatase A with UDP-N-Acetylglucosamine:Lysosomal Enzyme-N-Acetylglucosamine-1-phosphotransferase. *J. Biol. Chem.* 274(6):3651-3658.
24. Gieselmann V (1991) An assay for the rapid detection of the arylsulfatase A pseudodeficiency allele facilitates diagnosis and genetic counseling for metachromatic leukodystrophy. *Hum. Genet.* 86(3):251-255.
25. Ferguson RL, Pascreau G, Maller JL (2010) The cyclin A centrosomal localization sequence recruits MCM5 and Orc1 to regulate centrosome reduplication. *J. Cell Sci.* 123(16):2743-2749.
26. Pascreau G, Eckerdt F, Churchill MEA, Maller JL (2010) Discovery of a distinct domain in cyclin A sufficient for centrosomal localization independently of Cdk binding. *Proc. Natl. Acad. Sci. U.S.A.* 107(7):2932-2937.
27. Brown NR, et al. (1995) The Crystal-Structure of Cyclin-A. *Structure* 3(11):1235-1247.

28. Navin N, et al. (2010) Inferring tumor progression from genomic heterogeneity. *Genome Res.* 20(1):68-80.
29. Xu Z, et al. (2005) Liver-specific inactivation of the Nrf1 gene in adult mouse leads to nonalcoholic steatohepatitis and hepatic neoplasia. (Translated from eng) *Proc. Natl. Acad. Sci. U.S.A.* 102(11):4120-4125 (in eng).
30. Zent R, et al. (2010) Proteoglycans and Cancer. *Cell-Extracellular Matrix Interactions in Cancer*, (Springer New York), pp 191-215.
31. Sifaki M, et al. (2006) Lumican, a small leucine-rich proteoglycan substituted with keratan sulfate chains is expressed and secreted by human melanoma cells and not normal melanocytes. *IUBMB Life* 58(10):606-610.
32. Kato Y, et al. (2008) Increased expression of highly sulfated keratan sulfate synthesized in malignant astrocytic tumors. *Biochem. Biophy. Res. Commun.* 369(4):1041-1046.
33. Takahashi K, Stamenkovic I, Cutler M, Dasgupta A, Tanabe KK (1996) Keratan sulfate modification of CD44 modulates adhesion to hyaluronate. *J. Biol. Chem.* 271(16):9490-9496.
34. Thomas L, Byers HR, Vink J, Stamenkovic I (1992) Cd44h regulates tumor-cell migration on hyaluronate-coated substrate. *J. Cell Biol.* 118(4):971-977.
35. Malavaki C, Mizumoto S, Karamanos N, Sugahara K (2008) Recent advances in the structural study of functional chondroitin sulfate and dermatan sulfate in health and disease. *Connective Tissue Res.* 49(3-4):133-139.
36. Sasaki T, et al. (2006) *Balaenoptera omurai* is a newly discovered baleen whale that represents an ancient evolutionary lineage. *Mol. Phylogenet. Evol.* 41(1):40-52.
37. Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 92(2):532-536.
38. Xu XF, Arnason U (1996) A complete sequence of the mitochondrial genome of the Western lowland gorilla. *Mol. Phylogenet. Evol.* 13(5):691-698.
39. Bjornerfeldt S, Webster MT, Vila C (2006) Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res.* 16(8):990-994.
40. Arnason U, Gullberg A, Janke A, Kullberg M (2007) Mitogenomic analyses of caniform relationships. *Mol. Phylogenet. Evol.* 45(3):863-874.
41. Peng R, et al. (2007) The complete mitochondrial genome and phylogenetic analysis of the giant panda (*Ailuropoda melanoleuca*). *Gene* 397(1-2):76-83.
42. Sasaki T, et al. (2005) Mitochondrial phylogenetics and evolution of mysticete whales. *Syst. Biol.* 54(1):77-90.
43. Arnason U, Gullberg A, Janke A (2004) Mitogenomic analyses provide new insights into cetacean origin and evolution. *Gene* 333:27-34.
44. Gilbert MTP, et al. (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317(5846):1927-1930.
45. Poinar HN, et al. (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* 311(5759):392-394.
46. Krause J, et al. (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439(7077):724-727.
47. Rogaev EI, et al. (2006) Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *Plos Biol.* 4(3):403-410.

48. Arnason U, et al. (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci. U.S.A.* 99(12):8151-8156.
49. Douglas KC, et al. (2011) Complete mitochondrial DNA sequence analysis of Bison bison and bison-cattle hybrids: Function and phylogeny. *Mitochondrion* 11(1):166-175.
50. Pemberton D (1990) Social organization and behaviour of the Tasmanian devil, *Sarcophilus Harrisii*. Ph.D. (University of Tasmania, Hobart).
51. Lachish S, McCallum H, Jones M (2009) Demography, disease and the devil: life-history changes in a disease-affected population of Tasmanian devils (*Sarcophilus harrisii*). *J. Anim. Ecology* 78(2):427-436.