PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOLOGIA

História evolutiva e dinâmica demográfica de *Cavia intermedia* (Mammalia: Rodentia) inferida através de abordagem genômica

Manuel Adrian Riveros Escalona

Orientador: Prof. Dr. Sandro Luis Bonatto

DISSERTAÇÃO DE MESTRADO

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

Av. Ipiranga 6681 - Caixa Postal 1429

Fone: (051) 320-3500 - Fax: (051 339-1564

CEP 90619-900 Porto Alegre - RS

2015

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOLOGIA

História evolutiva e dinâmica demográfica de *Cavia intermedia* (Mammalia: Rodentia) inferida através de abordagem genômica

Manuel Adrian Riveros Escalona

Orientador: Prof. Dr. Sandro Luis Bonatto

DISSERTAÇÃO DE MESTRADO

PORTO ALEGRE – RS – BRASIL

2015

Sumário

Agradecimentos

Começarei dizendo que é difícil mencionar todas as pessoas que possibilitaram, direta ou indiretamente, a realização deste trabalho. Espero conseguir, neste pequeno espaço, prestigiar igualmente todos os envolvidos. Então, como prometido a um amigo, começo dizendo as palavras célebres de um encanador: "*Here we go!*"

Gostaria de agradecer primeiramente ao meu orientador Sandro Bonatto, tanto pela oportunidade de realizar este trabalho, como pela confiança para me permitir fazer algo que ninguém havia feito no laboratório.

Também não posso esquecer de mencionar a Fernanda Pedone, sem a qual este trabalho nunca teria saído do papel. Foram muitas horas discutindo o protocolo de laboratório e resolvendo os problemas que surgiam ao longo do caminho.

À Fernanda Trindade, nossa futura bioinformata, pelo auxilio na programação e a paciência para disponibilizar os computadores nessas últimas semanas. Acho que te devo um estoque vitalício de café.

Deixo aqui também um abraço ao Eckart, que viajou por muitos quilômetros para ajudar na bancada.

Os próximos agradecimentos se dirigem à família, e tenho muitas a quem agradecer.

Primeiramente à minha família de sangue, em especial à minha mãe, que sempre foi sinônimo de apoio e sacrifício para que eu pudesse chegar aonde estou. Este trabalho é pra ti!

À Letícia Frizzo, minha namorada que me suportou em todos os momentos e sempre esteve ao meu lado. Apesar dos trancos e barrancos, nós conseguimos! E é por isso que te amo!

Minha família genômica não pode faltar também. São muitas pessoas que iniciaram como colegas e conhecidos e se tornaram amigos e irmãos. Deixo um agradecimento especial para o Fernando, Biscoito (digo, Fabrício), Tiago Ferraz, Maísa, Isadora, Júlia e Lucas. Vocês tornaram a vivência (e sobrevivência) aqui mais fácil. Muitos verões escaldantes e invernos congelantes ainda virão para nós.

Agradeço à família da Sketch, em especial ao mestre Tanus, por todas as conversas sérias (e raras) e as besteiras, além do pouco desenho. Um forte abraço pro Anderson, Sabrina, Douglas, Luis, Renata, Thyago e Cristian. Veremos o Tanus ‘chorar’ de alegria ao ler isso...

Meus irmão de RPG (Bruno, Armando, Tito, Miguel e Nicolas), que sempre foram um cano de escape pros estresses da vida. Sempre teremos uma ilha!!! Se acharmos a Golden House, a primeira rodada é por minha conta. A segunda (com ninfas) é por conta do mago.

Minha família marcial também merece agradecimentos: Simone e Borghetti, que sempre serviram como mais que irmãos para mim. Ao Mestre Alexandre, um sinônimo de tranquilidade e caráter. E ao Roger, José, Daniel e Márcio pelas risadas e cervejas. Hapki e Haedong!!!

Aos amigos de graduação. Edna, Joice, Tahila, Letícia, Bruno, Nicolas, nunca teremos um nome para o grupo, mas sempre teremos as pizzas e os jogos de tabuleiro (e o queijo do jiban, o verme da pedra, e por ai vai).

Por último, mas não menos importante, aos amigos do Congen, com quem passei pouco tempo, mas valeu por uma vida. Adoraria chamar vocês como merecem, mas não seria apropriado nesse texto. Matt, Wilfried, Nicole, Jane e Inger, no próximo aniversário em Puerto Rico a primeira rodada de pitorro é por minha conta (e possivelmente a segunda).

Finalizo com um agradecimento aos que, pela minha falta de memória, não foram mencionados: “Até mais e obrigado pelos peixes”.

Resumo

Populações insulares sempre tem sido objeto de interesse para estudos genéticos e ecológicos. A espécie insular *Cavia intermedia* é um dos mamíferos mais raros no mundo, com um tamanho de censo populacional de aproximadamente 40 indivíduos, corroborado por uma diversidade genética e tamanho populacional efetivo (N_e) extremamente baixos. Aqui, nós usamos sequenciamento de DNA associado a sítios de restrição (no inglês, RAD-seq) para obter dados de sequencias ao longo do genoma de *C. intermedia* e sua espécie-irmã continental, *C. magna*, para estudar sua diversidade molecular e história demográfica. Nós geramos um total de 10 milhões de *reads* de 300 bp para 20 indivíduos de cada espécie. Entretanto, mesmo usando o genoma de *Cavia porcellus* como referência ou abordagem *de novo* e aplicando filtros rigoroso para qualidade e dados faltantes, apenas uma centena de *loci* puderam ser usados. Este conjunto final de dados gerou valores de diversidade molecular e tempo de divergência amplamente diferentes entre os métodos de análise, alguns compatíveis com estudos anteriores (e.g., N_e ~60 para *C. intermedia*) mas alguns com valores significativamente maiores. Nós sugerimos que para obter um conjunto de dados de alta qualidade e muito informativo seria necessário aumentar significativamente os dados de sequência bem como aplicar outras abordagens de montagem.

Abstract

Insular populations have always been the subject of interest for ecological and genetic studies. The insular *Cavia intermedia* is one of the rarest mammals in the world, with an average census size of about 40 individuals, corroborated by an extremely low genetic diversity and effective population size (N_e). Here, we used restricted site-associated DNA sequencing (RAD-seq) to obtain genome wide sequence data of *C. intermedia* and its continental sister species, *C. magna* to study their molecular diversity and demographic history. We have generated approximately 10 millions of useable reads of 300 bp for 20 individuals each of both species. However, both using the *Cavia porcellus* draft genome as a reference or *de novo* approaches and applying stringent filters for quality and missing data, only about a hundred loci could be used. This final dataset generated widely different molecular diversity and divergence time values between methods of analyses, some compatible with the previous studies (e.g., Ne ~60 for *C. intermedia*) but some with significantly larger values. We suggest that to obtain a high quality and very informative dataset it would be necessary to significantly increase the sequence data as well as apply other assembly approaches.

Apresentação

Cavia intermedia é uma espécie endêmica das ilhas de Moleques do Sul, em Santa Catarina, descrita recentemente e que tem chamado a atenção de diversos pesquisadores. A espécie é observada em apenas uma ilha de 10ha a 14 km da costa de Santa Catarina (Cherem *et al.* 1999), ocupando um pequeno fragmento da ilha e apresenta um tamanho populacional censitário extremamente reduzido, entre 30 e 60 indivíduos (Salvador & Fernandez 2008a; b). Acredita-se que tenha se separado de *Cavia magna*, espécie com ampla distribuição desde o Uruguai até Santa Catarina e considerada a espécie-irmã (Cherem *et al.* 1999; Kanitz 2009; Furnari 2013), no mínimo por volta de 8.000 anos atrás, época em que o aumento do nível do mar após a última glaciação cortou a conexão terrestre entre a ilha e o continente (Corrêa 1996; Cherem *et al.* 1999).

Esse longo isolamento geográfico faz com que *C. intermedia* apresente diversas das características observadas em populações insulares em relação às contrapartes continentais: ninhadas reduzidas, constituídas de 1 ou 2 indivíduos, menor tamanho e longevidade (Salvador 2006; Salvador & Fernandez 2008a; b). Soma-se às diferenças morfo-fisiológicas o fato da espécie apresentar um cariótipo distinto das demais espécies do gênero *Cavia* (Gava *et al.* 1998) e uma menor diversidade genética.

Kanitz (2009), usando marcadores mitocondriais e 12 loci de microssatélites nucleares, demonstrou que a espécie apresenta uma diversidade muito reduzida. Quando comparada com *C. magna*, *C. intermedia* apresenta uma diversidade sete vezes menor para os *loci* de microssatélite. Os resultados ainda corroboram as observações de Salvador & Fernandez (2008a), sugerindo que a espécie vem mantendo uma população muito pequena de maneira relativamente constante por muitas gerações, sem efeito *bottleneck* no passado recente. Ademais, simulações realizadas por Salvador (2006) demonstraram probabilidade de 100% da espécie se extinguir nos próximos 100 anos, com tempo médio de extinção de 22 anos.

Esse conjunto de características torna *C. intermedia* uma das espécies mais raras do planeta (Salvador 2006; Salvador & Fernandez 2008a; b; Kanitz 2009) e um importante objeto de estudo para a compreensão dos processos evolutivos que atuam sobre pequenas populações.

O uso de métodos multilocus para inferir história evolutiva, demografia, filogenia e filogeografia de diferentes populações se tornou a pedra-fundamental dos estudos

moleculares nos últimos trinta anos (McCormack *et al.* 2013). Isso foi potencializado pelo desenvolvimento de técnicas de sequenciamento automatizado, dentre os quais a técnica de Sanger tem sido amplamente utilizada, sendo limitada apenas pelo número de amostras sequenciadas simultaneamente (Shendure & Ji 2008).

O advento do sequenciamento de nova geração (*next-generation sequencing*; NGS) a partir de 2005 permitiu a geração de uma grande quantidade de dados à um custo progressivamente mais baixo, gerando uma "democratização" para o sequenciamento de organismos (Ekblom & Galindo 2011). O aumento expressivo no número de marcadores possibilita maior precisão em estudos de genética de população e nos padrões demográficos (Duran *et al.* 2009; Hoglund 2009; Ekblom & Galindo 2011).

Apesar de suas vantagens, o uso dessas novas tecnologias em estudos filogenéticos e filogeográficos tem sido lento devido ao objeto de estudo nessas áreas: geralmente organismos não-modelo (*non-model organism*), onde pouca ou nenhuma informação genômica prévia da espécie ou de espécies próximas existe. Essa falta de informação somada a necessidade de se sequenciar grande quantidade de indivíduos torna os métodos tradicionais de NGS, como o sequenciamento total do genoma (*whole-genome sequencing*; WGS) inviáveis, tanto pelo custo como pelo esforço analítico (McCormack *et al.* 2012, 2013).

Diversos métodos foram desenvolvidos para se trabalhar com muitos indivíduos a um custo razoável, consistindo principalmente na simplificação da biblioteca através do sequenciamento de porções do genoma, sejam estas aleatórias ou previamente selecionadas, gerando milhares de marcadores (Ekblom & Galindo 2011; McCormack *et al.* 2012, 2013; Andrews & Luikart 2014). Dentre os métodos existentes, o conjunto de protocolos conhecido como RADseq (*Restriction-site-Associated DNA sequencing*) vem sendo empregado de maneira eficiente em diversos trabalhos (Rowe *et al.* 2011). O protocolo original foi proposto por Baird *et al.* (2008) e diversas adaptações surgiram, dentre elas ddRADseq (Peterson *et al.* 2012) e RESTseq (Stolle & Moritz 2013). Nestes métodos, enzimas de restrição são usadas para fragmentar o genoma e selecionar apenas parte dos fragmentos, resultando em uma redução no custo da geração da biblioteca. Adicionalmente, evita-se a necessidade de grandes volumes de material, permitindo o sequenciamento com volumes menores de DNA (100ng ou menos). Os protocolos também permitem a seleção de fragmentos de acordo com o tamanho, permitindo um melhor controle das regiões representadas na biblioteca.

Esse conjunto de características permite a geração de bibliotecas com fragmentos derivados de centenas a centenas de milhares de regiões do genoma com apenas pequenas porções de material. Para espécies ameaçadas, como *C. interemedia,* onde a quantidade de material é limitada e a coleta deve ser não-invasiva, este se mostra um método ideal para avaliação do genoma (Dale & von Schantz 2002).

O uso dessas novas tecnologias, entretanto, gerou uma nova gama de desafios que devem ser superados para possibilitar a utilização adequada desses dados (Nielsen *et al.* 2011; Yu & Sun 2013; Bragg *et al.* 2013; Ross *et al.* 2013; Davey *et al.* 2013). A primeira dificuldade a ser superada é a geração dos *contigs* para a detecção de polimorfismos. Apesar de alguns genomas existirem, possibilitando o alinhamento a uma referência, a maioria dos táxons de interesse para a conservação não possuem um genoma de referência de boa qualidade. Dessa forma, duas abordagens costumam ser realizadas. A primeira consiste em usar um genoma uma espécie próxima como molde. Apesar de eficiente, esta técnica apresenta muitos vieses para estudos intraespecíficos, devido à possibilidade polimorfismos únicos da espécie serem identificados. A segunda opção amplamente utilizada é a geração dos *contigs* através de uma montagem *de novo,* onde os próprios fragmentos gerados servem como base para a formação de sequências maiores. O uso de algoritmos avançados torna a montagem *de novo* mais apropriada para estudos intraespecíficos, mas a falta de uma referência pode resultar no agrupamento de regiões parálogas e na superestimativa de polimorfismos (Li 2011; Amores *et al.* 2011; Eaton 2014).

Além da geração dos *contigs*, a natureza aleatória do sequenciamento NGS resulta em diferentes coberturas ao longo do genoma. Isso faz com que, mesmo utilizando técnicas como RAD-seq, muitos conjuntos de dados apresentam coberturas que variam de 1X a 1000X. Duas principais consequências disso são a incapacidade de determinar o genótipo do indivíduo, devido às incertezas associadas à baixa cobertura, e a alta proporção de dados faltantes, seja pelo não sequenciamento da região ou pelos filtros empregados para melhorar a qualidade do conjunto de dados. Dessa forma, é comum nesse tipo de estudos se observar o descarte de grandes porções do conjunto de dados, o que também pode gerar mais viés nos resultados obtidos (Huang & Knowles 2014).

O objetivo deste trabalho foi utilizar tecnologia NGS associada à técnica RAD-seq para gerar um grande volume de dados para auxiliar na elucidação de questões referentes à diversidade genética de *Cavia intermedia* e seu histórico demográfico,

comparando os resultados com sua espécie-irmã, *Cavia magna*. Os resultados serão apresentados sob a forma de um *Original Article* no formato do periódico *Molecular Ecology*.

Original Article

Evolutionary history and population dynamics of *Cavia intermedia* inferred with a RAD-seq approach

Manuel A. R. Escalona[1], Sandro L. Bonatto[1]

[1] Faculdade de Biociências, PUCRS, Av. Ipiranga 6681, prédio 12C. Porto Alegre, RS 90619-900, Brazil

Keywords: island endemic species, NGS, demographic estimates, divergence time

Corresponding author: Sandro L. Bonatto, Faculdade de Biociências, PUCRS. Av. Ipiranga 6681, prédio 12C, sala 134. Porto Alegre, RS 90619-900, Brazil. E-mail: slbonatto@pucrs.br

Abstract

Insular populations have always been the subject of interest for ecological and genetic studies. The insular *Cavia intermedia* is one of the rarest mammals in the world, with an average census size of about 40 individuals, corroborated by an extremely low genetic diversity and effective population size (N_e). Here, we used restricted site-associated DNA sequencing (RAD-seq) to obtain genome wide sequence data of *C. intermedia* and its continental sister species, *C. magna* to study their molecular diversity and demographic history. We have generated approximately 10 millions of useable reads of 300 bp for 20 individuals each of both species. However, both using the *Cavia porcellus* draft genome as a reference or *de novo* approaches and applying stringent filters for quality and missing data, only about a hundred loci could be used. This final dataset generated widely different molecular diversity and divergence time values between methods of analyses, some compatible with the previous studies (e.g., Ne ~60 for *C. intermedia*) but some with significantly larger values. We suggest that to obtain a high quality and very informative dataset it would be necessary to significantly increase the sequence data as well as apply other assembly approaches.

Introduction

Insular populations have always been of great interest to ecological and conservation studies. The viable colonization of islands by continental species, be it by geographic isolation or migration, results in a series of morphological, physiological and genetic modifications (Adler & Levins 1994; Adler 1996; Blanco *et al.* 2014). Among these modifications, behavioral and feeding innovations are associated to adaptations to the isolated, depleted and unique environments present on islands, being facilitated by the reduction of interspecific competition and predation. Evidence has also shown that these new populations persist when able to enter in the new food chains, exploring new, non-optimal and even toxic resources (Blanco *et al.* 2014).

Analyzing demographic patterns, insular populations tend to present a reduced population compared to their continental counterparts (Adler & Levins 1994). This reduced population size (N) directly affects the genetic diversity of given population, reducing the effective population size (N_e), this being responsible for the maintenance of population genetic variability (Frankham *et al.* 2002; Begon *et al.* 2006). N_e depends on

different factors, mainly reproductive success, sex ratio, population size fluctuation and stochastic events (Reed 2010).

N_e can be calculated considering an idealized population in Hardy-Weinberg equilibrium (Hoglund 2009). However, real biological systems do not follow idealized patterns and usually present a lower than expected N_e that is also much lower than the observed census population (Palstra & Ruzzante 2008). N_e reduction and, consequently loss of evolutionary potential is considered a main cause of population decline and species extinction (Frankham *et al.* 2002; Begon *et al.* 2006). Low diversity added to ecological and demographic stochastic factors make insular species highly susceptible to extinction.

Among cases cited in literature, we can mention the work of Seddon & Baverstock (1999) with native australian rats (*Ratus fuscipes greyii*) dispersed in 15 islands in the south coast of Australia. Aside from the correlation between genetic diversity and the distance to the continent and island area, they suggest that part of the low diversity is due to successive population bottlenecks. In another research with insular platypus (*Ornithorhynchus anatinus*), Furlan *et al.* (2012) also found low diversity values, added to high inbreeding. Grazziotin *et al.* (2006), studying two insular species from the *Bothrops* genus in Brazilian Atlantic rainforest found similar patterns for mitochondrial markers.

There is intense debate about the minimum viable population size required to avoid the extinction of species. Genetically viable populations are the ones capable of avoiding inbreeding depression and, therefore, the accumulation of deleterious mutations, keeping the evolutionary potential (Traill *et al.* 2010). Estimates suggest a N_e around 500 individuals for a viable population, with ~50 individuals as a minimum to avoid inbreeding (Frankham 1995). Reed (2010) suggests that a species persistence time (*i.e.* capacity of not going extinct) is the result of several factors, like environmental stress, genetic diversity and biological characteristics. Therefore, if a population is adapted to a stable environment (*i.e.* niche availability, few pathogens and/or predators, stable climate, etc.), low genetic variability will hardly be a limiting factor, making it difficult to estimate how low genetic diversity can be and still not be a cause of extinction (Reed 2010).

Although there are methods to more directly estimate the viability of a population, most of them were developed based on studies under controlled conditions, in captive or reintroduced species or even through mathematical models (Begon *et al.* 2006).

Additionally, the species with very small population size studied so far have been mainly reduced through anthropic action (Begon *et al.* 2006; IUCN 2012; Animal Info 2014), with few studies in populations that are naturally small.

In this context, the Moleques do Sul cavy *Cavia intermedia* presents itself as an ideal model for conservation genetics studies. The species is only observed in a 10ha island 14 km from the state of Santa Catarina, Brazil (Cherem *et al.* 1999) (Fig.1) and is believed to have diverged from the continental guinea pig (*Cavia magna*), with distribution from Uruguay to Santa Catarina, at least 8 000 YA, when the rising sea level last separated the land between island and continent (Corrêa 1996; Cherem *et al.* 1999; Kanitz 2009; Furnari 2013).

Since its description by Cherem *et al.* (1999) some studies have been made with *C. intermedia*, evaluating population dynamics, social behavior, karyotype and evolutionary history (Gava *et al.* 1998; Salvador & Fernandez 2008a; b; Kanitz 2009; Furnari 2011, 2013). Very interestingly, Salvador & Fernandez (2008a, b) found that, due to the island very small area and available habitat, the whole species census size during the 15 months of study fluctuated between around 30 and 60 individuals, with an average of 42 individuals. Furthermore, simulations made by Salvador (2006) showed a 100% probability of extinction in the next 100 years, with an average time to extinction of 22 years.

Kanitz (2009), using mitochondrial markers and 12 nuclear microsatellite loci showed that *C. intermedia* presents a very reduced genetic diversity. When compared to *C. magna*, *C. intermedia* is seven-times less diverse for the STR loci and presents a single mtDNA haplotype. However, this *C. intermedia* mtDNA haplotype is very different from the *C. magna* haplotypes, with an estimated divergence time of more than a million years (Kanitz 2009). This genetic study also found the species presents a N_e of around 42 individuals with no signal of recent population reduction, corroborating the low census size observed by Salvador & Fernandez (2008b).

This group of characteristics make *C. intermedia* one of the rarest species on the planet (Salvador 2006; Salvador & Fernandez 2008a; b; Kanitz 2009) and an important study subject for the understanding of evolutionary processes that act on small populations.

In this paper we re-evaluate the genetic variability and evolutionary history of *Cavia intermedia* through NGS techniques, specifically using a RAD-seq approach.

Material and Methods

Sampling

Twenty individuals from each species were used in the generation of RADseq data. Ear tips from *Cavia intermedia* were obtained during the collection made by Salvador & Fernandez (2008) between May 2004 and June 2005 in the Moleques do Sul Archipelago and conserved in alcohol 70% at -80ºC. Tissue samples from *Cavia magna* were collected in 2013 at the district of Pinheira in the Santa Catarina coast, in front of the archipelago, and maintained under the same conditions.

Extraction and RADseq Library Preparation

DNA was extracted using QIAGEN Blood and Tissue kits, followed by quality verification in 1% agarose gel and quantification using L-Quant (Loccus Biotecnologia). For each sample, we made two elutions and the one with enhanced DNA quality was used for library preparation.

Library preparation followed the RESTseq protocol (Stolle & Moritz, 2013), using 1μg of genomic DNA. The protocol, designed for Ion Torrent, consists of a first digestion using a common-cutting enzyme followed by adapter ligation and purification followed by a second digestion. For the second digestion, a rare-cutting enzyme or a mix of enzymes is used for library complexity reduction.

The first digestion reaction was incubated for 5 hours with 100U of TaqI enzyme (Promega), followed by barcoded adapter ligation. We selected only barcodes without any enzyme recognition site to avoid library digestion. The samples went through a second digestion in a mix containing 100U of HaeIII and MboI for complexity reduction. Samples were then size-selected to approximately 350bp with E-Gel (Life Technologies). Samples were pooled and quantified through qPCR and sequenced using the PGM Ion Torrent platform along with the 400bp sequencing kit and Life Technologies 318 chip.

The choice of enzymes was made through the SimRAD (Lepais & Weir 2014) package for R (R Core Team 2014), using the unmasked *Cavia porcellus* (cavPor3 from

ensemble.org) genome as reference. We aimed to sequence around 8 000 loci per individual to maximize coverage per locus. Simulation of the total number of fragments used 50% of the whole genome to estimate RAD sites, due to computational memory limitation, and considered an average depth of 30X and fragment length ranging from 320bp to 380 bp, resulting in approximately 8 500 loci per individual.

Bioinformatics

We used two approaches to evaluate genomic data. The first consisted in assembling a SNP matrix using the software Samtools (Li 2011), while for the second we used sequences obtained with the software STACKS (Catchen *et al.* 2011, 2013). Both approaches went through the same initial filtering and alignment and were used for analyses with both species together or with each species separately.

Ion Torrent Ion Suite software demultiplex samples prior to exporting them to fastq format. The software also filters for polyclonal and primer dimer reads and reports low quality reads. Raw reads were trimmed to 200bp in order to obtain the maximum dataset while still removing smaller reads and possible bias during contig assembly. Reads were then filtered to remove sequences where the average quality dropped below 99% probability of being correct in a 30bp window using the software *process_radtags* provided with STACKS. After this initial step, reads were aligned in Bowtie2 (Langmead & Salzberg 2012) to the masked *Cavia porcellus* genome in order to filter for repetitive regions. The masked genome was obtained from ensembl database (ensemble.org) and consists on the same scaffolds from cavPor3 with known repetitive regions replaced by "N". This results in some reads not aligning and being dropped (*i.e.* discarded) by Bowtie2. Repetitive regions not present in the database could pass this filtering step and result in bias and were, whenever possible, manually removed from the dataset when generating input files for analysis.

In STACKS, we used the program *ref_map.pl* to group the aligned reads based on the alignment, setting a minimum depth of 5X as a threshold. The result is a catalog of loci (also called RAD-tags) spread throughout the genome with the observed genotypes, consensus sequences and identified SNPs. The catalogs went through a correction step using the recently implemented program *rxstacks.pl* with the options –prune_haplo and –conf_lim 0.25 for removal of potential paralogous, sequencing errors and miscalled polymorphisms. Fragments generated in STACKS (from here on called

RAD-tags) were exported in FASTA through *populations* software, with a minimum 10X coverage and allowing for 20% of missing data. For comparative data between species, it was selected only loci present in both species. Exported sequences were filtered to remove loci with 4 or more clustered SNPs in a 50 bp window and with too high polymorphism (possibly from sequencing error and repetitive or paralogous regions that passed the alignment) (Frantz *et al.* 2014) through FilterVariation, available in GATK (McKenna *et al.* 2010).

The SNP matrix (henceforward called SNP-tags) was built using Samtools 1.1 *mpileup* command. We used *vcfutils.pl* to filter the data, keeping only SNPs with coverage above 30X and with a minimum 10bp distance from gaps. SNP clusters were filtered using the same method described for exported sequences. SNP-tags were kept in VCF format, allowing for up to 50% of missing data (Huang & Knowles 2014), followed by more stringent parameters according to the analysis.

The software PGDSpider2 (Lischer & Excoffier 2012) was used transform fasta or VCF to the input files for the different analyses.

Demographic and Divergence Time Estimation

Summary statistics for the datasets were calculated with the software Arlequin (Excoffier & Lischer 2010), while the R package PopGenome (Pfeifer *et al.* 2014; R Core Team 2014) was also used for SNP-tags. For most analyses, a mutation rate (μ) of 4.9 X 10^{-8} per generation (Lischer *et al.* 2014) was used.

RAD-tags were used as independent loci to estimate demographic fluctuations through time with the *Extended Bayesian Skyline Plot* method (Heled & Drummond 2008) implemented in BEAST2 (Bouckaert *et al.* 2014). For this analysis we used a strict molecular clock with the above mutation rate, HKY substitution model and 1 billion iterations, discarding the first 100 million iterations as burn-in.

Migrate-n (Beerli 2006) Bayesian approach was also used to estimate the theta parameter ($4N_e\mu$) and demographic history using its built-in skyline approach. After some initial runs to evaluate parameter limits, a final run was made using a theta multiplication prior with mean 5 X 10^{-5}, maximum 0.01, delta 100 and 5 000 bins. Sampling occurred every 20 steps to a total of 40 000 samples with 100 000 samples discarded as burn-in. We also used the Bayesian approach in Lamarc (Kuhner 2006) for estimation of theta and

growth rate, using the same prior information used on Migrate-n analysis. To transform the theta ($4N_e\mu$) value to effective population size (N_e) we used the above μ and in the case of skyline time estimates and growth rates, a generation time of 0.33 (Salvador & Fernandez 2008b).

Using both species dataset (*C. intermedia* and *C. magna*), the BPP (Rannala & Yang 2003; Yang & Rannala 2010) and GPhoCS (Gronau *et al.* 2011) sofywares were also used to estimate the divergence time of the species and the effective population sizes. BPP accommodates the species phylogeny as well as incomplete lineage sorting due to ancestral polymorphism. GPhoCS is inspired in MCMCcoal to infer ancestral population sizes, divergence times and migration rates from individual genome sequences or separate loci along the genome. In both analysis we set a theta prior G(2, 500 000) for both species and a tau (divergence time) prior G(10, 500) for the age of the root. These analyses were run twice to confirm the obtained values.

For SNP-tags we used the software MSMC (Schiffels & Durbin 2014) to estimate the demographic history for each species separately. This software works similar to PSMC (Li & Durbin 2011), allowing to infer population size and gene flow from a single or multiple genomes. For this analysis, one individual from each species was randomly chosen for an initial run, followed by the addition of individuals as computer processing capacity allowed.

Results

Sequencing and Alignment

We prepared 4 libraries to a total of 17 824 104 reads with size ranging from 270 to 330 bp (Table 1) in four runs (chips). From these, 6 849 068 (ranging between 62 389 and 316 173 by sample) were retained after the initial filtering in *process_radtags*.

The alignment resulted in an average of 54.98% of reads not aligned (minimum and maximum of 47.69% and 59.52%, respectively), 10.46% (7.95% - 21.37%) aligned in one place and 34.56% (30.41% - 37.82%) in more than one place along the masked genome (Supplementary Table S1).

SNP detection and validation

Initial STACKS catalogs for each species consisted of approximately 15 000 loci, while the catalog for both species contained 9 390 loci. Filtering of putative paralogous

and sequencing error through *rxstacks.pl* resulted between 58 regions with high coverage and presence of individuals for both species, 71 loci for *C. intermedia* and 88 loci for *C. magna* for posterior analysis, with 184 and 205 polymorphic sites, respectively (Table 2). Almost all loci discarded during this step were due to the amount of missing data. Similar results were observed for SNP-tags (Table 3).

SNP-tags generated in Samtools resulted in an average of 1 023 897 variants prior to filtering. The high quality matrix consisted of approximately 30 000 SNPs for each species dataset and 98 000 for both species combined, with the mean depth of coverage being 43X. The proportion of missing data within species was of 30%, raising to around 65% when both species were combined (Table 3). This high amount of missing data resulted in the impossibility to generate the required inputs for MSMC analysis, making SNP-tags data unusable.

Demographic History and Divergence Time

Summary statistics resulted in over 10 thousand nucleotides for RAD-tags and over 29 thousand variable sites for SNP-tag data. RAD-tags showed similar heterozygosity between species, while SNP-tags showed a heretozygosity 4 times higher for *C. magna* (Table 3).

Our estimates for the current effective population size resulted in very divergent values between the different methods used (Table 4). BPP predictions resulted in the lowest values among all analyses for the current effective size for *C. intermedia* and *C. magna* (62 and 91 individuals respectively). G-PhoCS analysis resulted in the high effective population size values, being 46 719 individuals for *C. intermedia* and 75 204 individuals for *C. magna*. We focused the other estimates in *C. intermedia* only. Migrate estimates suggests a *C. intermedia* effective population of size 17 245 individuals, while Lamarc resulted in an estimate of 11 010 individuals (Table 4). Ancestral population effective sizes ranged from 2 452 individuals for BPP analysis to as high as 25 341 827 individuals in G-PhoCS.

Migrate *Skyline Plot* for *C. intermedia* suggests a decrease in effective population size until approximately 200 years ago, when supposedly the population started to increase again. (Fig.2).

Divergence time was estimated to be around 143 414 years, with 95%HPD ranging from 53 185 and 235 622 using G-PhoCS. BPP divergence time estimates resulted in a dating of 56 years.

Discussion

In this study we tried to generate a high quantity of sequence data through RAD-seq with the objective of estimating the current and historical effective population size of *Cavia intermedia* and the divergence time from its continental sister species, *Cavia magna*. Most methods to estimate demographic history from NGS depends on a high amount of SNP data and/or a high quality reference genome (Sheehan *et al.* 2013; Excoffier *et al.* 2013; Schiffels & Durbin 2014), although the use of *de novo* assembly methods (Atwood *et al.* 2001; Catchen *et al.* 2011; Eaton 2014) have also become a viable option when no prior information is available. The RAD-seq method has been well documented (Baird *et al.* 2008) and widely used for non-model species with no reference genome or sequence variation information (Davey *et al.* 2013), making it a good candidate to be useful for our study.

Sequencing and Genotyping Variation and SNP detection

No sequencing platform is free of errors and biases must be taken into account when designing the project and analyzing the results (Bragg *et al.* 2013; Ross *et al.* 2013). Aside from the platform specific potential problems, RAD techniques also have been known to present problems, such as great variation in read depths whose causes are not clear yet and therefore are very difficult to anticipate (Nielsen *et al.* 2011; Yu & Sun 2013; Davey *et al.* 2013). Even though the use of a reference genome to estimate the number of RAD loci helps improve depth of coverage, our alignments showed that, along with sequencing stochasticity, there is a high amount of repetitive regions being sequenced resulting in a great amount of discarded data during SNP and sequence filtering (see Table 3). *C. porcellus* average genome coverage of 7X might result in additional bias during alignment and SNP calling due to errors present in the actual genome assembly.

The filtering and proportion of missing data is another subject of discussion. More conservative approaches would require the removal of any loci with missing genotype and/or low coverage. Although there are methods to correct for low coverage (*e.g.* GATK Unified Genotyper), a very conservative approach has been shown to bias NGS analysis (Huang & Knowles 2014).

Given the availability of a draft genome from *Cavia porcellus* (domestic guinea pig), our first approach was to use it as reference genome in the assembling steps. However, it seems that several issues interfered with the output of good results with this approach. First, *C. porcellus* is a draft genome of not high quality and not particularly well annotated. Second, even using a genome masked for repetitive regions, we still found in our final contigs regions that were identified as repeats in the annotations or even of mitochondrial origin that should have been masked. Third, *C. porcellus* is not closely related to our focal species, being estimated to have diverged nearly 6 MYA (Dunnum & Salazar-Bravo 2010) and presents a different karyotype (Gava *et al.* 2011). Therefore, too stringent alignment parameters increases the proportion of non-aligned reads simply due to the divergence between the genomes while less stringent parameters increases the proportion of paralogous loci.

However, the alternative *de novo* assembly approach may also result in an excess of called SNPs due to the grouping of paralogous and repetitive regions. Since version 1.21, STACKS (Catchen *et al.* 2011, 2013) has a correction algorithm implemented to correct generated catalogs and filter tags according to different parameters. Still, since *de novo* assemblers take into account sequence similarity together with maximum likelihood calculations, no method exists to specify the best assembly parameters (Catchen *et al.* 2011; Eaton 2014; Mastretta-Yanes *et al.* 2014). As a result, a test run using STACKS *denovo_map.pl* and different assembly parameters resulted in catalogs ranging from 30 000 to 84 000 different loci (against the 8 000 simulated and 15 000 assembled loci).

The filtering steps resulted in a major loss of data for both RAD-tags and SNP-tags datasets. Although there can be many reasons for that, from sequencing error to probabilistic uncertainty from the method used to call SNPs, we believe that one of the major factors influencing our dataset was the sequencing randomness. Even being sequenced in the same chip and having the same amount of DNA input through the whole library preparation steps, many individuals showed different number of reads. This influence directly the mean depth of coverage for each individual, resulting in some regions being highly covered for some individuals and poorly covered for others (Nielsen *et al.* 2011). Simulations using SimRAD did not considered these factors.

Variations in depth of coverage resulted in many SNPs not being called for some individuals, causing a high proportion of missing data, in either RAD-tags or SNP-tags. Although there is no consensus as to how much missing data can be allowed in analyses,

too much missing information will result in large uncertainty estimates (Arnold *et al.* 2013; Huang & Knowles 2014).

Our results suggest that to obtain a reasonably large number of high-quality loci to be usable in demographic studies in species with large genomes, such as mammals, would require a very high sequence throughput, much higher than the one obtained here.

Molecular diversity and demographic parameters

Estimates of the theta population genetic parameter and consequently the current effective population size for *Cavia intermedia* and *C. magna* varied greatly among the methods used and mostly provided values much higher than the census size (Salvador & Fernandez 2008a) and the ones estimated previously using mtDNA and microsatellite data (Kanitz 2009). Divergence times between *Cavia intermedia* and *C. magna* also varied widely, from as low as 55 years to about 140 thousand years, the latter being still much lower than the more than a million years estimated previously from the mtDNA data (Kanitz 2009).

This large difference between our estimates by itself suggests that, even after extensive filtering for high-quality information, many biases remained in the datasets. Increasing the stringency of the filter parameters would result in too few loci to allow reasonable estimates of diversity values.

Interestingly, in all our diversity and N_e estimates, our *C. magna* population presented values very similar to those of *C. intermedia*, while in our previous estimates with microsatellites its genetic diversity was at least 7-times higher (Kanitz 2009). Even considering our dataset may still presents errors such as paralogous loci, this previous results suggest a higher number of called SNPs should be expected in *C. magna*. However, these previous estimates used individuals from different populations some hundreds of kilometers apart, while the present sample is from a single and small population that may have suffered anthropic reduction. Further analyses are required to test this hypothesis.

However, our extreme cautiousness with the above results is not only because they showed much higher than expected values, but mainly due the inconsistent results between the methods. Discrepancies between results obtained from NGS and traditional studies with just a single or a few loci are not unexpected (e.g. Jones *et al.* 2013), since

RAD-seq and similar approaches present genome wide data, being able to sequence regions that are either neutral or under selection (Xu *et al.* 2014).

In summary, although the RAD-seq approach we used generated a large quantity of raw sequence reads, the lack of a high quality reference genome, the large genome with a large proportion of repetitive regions, the relatively lower throughput of our equipment, coupled with the large uncertainties about the values of the parameters in the many different RAD-seq pipelines in use today; all these issues so far prevented us to obtain a fully trustful dataset. We are currently exploring other approaches to deal with the above issues in our data to generate a high quality dataset to help us to better understand the evolutionary history of *Cavia intermedia*.

References

Adler GH (1996) The island syndrome in isolated populations of a tropical forest rodent. *Oecologia*, **108**, 694–700.

Adler GH, Levins R (1994) The Island Syndrome in Rodent Populations. *The quaterly Review of Biology*, **69**, 473–490.

Animal Info (2014) Animal Info. - Endangered animal of the world.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology*, **22**, 3179–90.

Atwood TS, Gribbin JM, Boone JQ *et al.* (2001) RAD LongRead : a SNP Discovery and de novo Sequence Assembly Strategy. , 97403.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Beerli P (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–5.

Begon M, Townsend CR, Harper JL (2006) *Ecology: from individuals to Ecosystems*. Blackwell Publishing.

Blanco G, Laiolo P, Fargallo J a. (2014) Linking environmental stress, feeding-shifts and the "island syndrome": A nutritional challenge hypothesis. *Population Ecology*, **56**, 203–216.

Bouckaert R, Heled J, Kühnert D *et al.* (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis (A Prlic, Ed,). *PLoS Computational Biology*, **10**, e1003537.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Computational Biology*, **9**, e1003031.

Catchen JM, Amores A, Hohenlohe P *et al.* (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *Genes|Genomes|Genetics*, **1**, 171–182.

Catchen J, Hohenlohe P a, Bassham S, Amores A, Cresko W a (2013) Stacks: An analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Cherem JJ, Olimpio J, Ximenez A (1999) Descrição de uma espécie do Gênero Cavia Pallas, 1766 (Mammalia - Caviidae) das Ilhas dos Moleques do Sul, Santa Catarina, Sul do Brasil. *Biotemas*, **12**, 95–117.

Corrêa ICS (1996) Les variations du niveau de la mer durant les derniers 17.500 ans BP : l ' exemple de la plate-forme continentale du Rio Grande do Sul - Brésil. *Marine Geology*, **130**, 163–178.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Dunnum JL, Salazar-Bravo J (2010) Molecular systematics, taxonomy and biogeography of the genus Cavia (Rodentia: Caviidae). *Journal of Zoological Systematics and Evolutionary Research*, **48**, 376–388.

Eaton D a R (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, **9**, e1003905.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–7.

Frankham R (1995) Conservation genetics. *Annual review of genetics*, **29**, 305–27.

Frankham R, Ballou JD, Briscoe DA (2002) *Introduction to Conservation Genetics*. Cambridge University Press, New York, NY.

Frantz L a F, Madsen O, Megens H-J, Groenen M a M, Lohse K (2014) Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island Southeast Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Molecular ecology*, 5566–5574.

Furlan E, Stoklosa J, Griffiths J *et al.* (2012) Small population size and extremely low levels of genetic diversity in island populations of the platypus, Ornithorhynchus anatinus. *Ecology and Evolution*, **2**, 844–857.

Furnari N (2011) Comportamento e organização social do preá Cavia intermedia, uma espécie endêmica das Ilhas Moleques do Sul, Santa Catarina. Universidade de São Paulo.

Furnari N (2013) New findings on the origin of Cavia intermedia, one of the world's rarest mammals. *Mammal Review*, **43**, 323–326.

Gava A, Freitas TRO, Olimpio J (1998) A new karyotype for the genus Cavia from a southern island of Brazil (Rodentia - Caviidae). *Genetics and Molecular Biology*, **21**, 77–80.

Gava A, Santos MB, Quintela FM (2011) A new karyotype for Cavia magna (Rodentia: Caviidae) from an estuarine island and C. aperea from adjacent mainland. *Acta Theriologica*, **57**, 9–14.

Grazziotin FG, Monzel M, Echeverrigaray S, Bonatto SL (2006) Phylogeography of the Bothrops jararaca complex (Serpentes: Viperidae): Past fragmentation and island colonization in the Brazilian Atlantic Forest. *Molecular Ecology*, **15**, 3969–3982.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, **43**, 1031–1034.

Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC evolutionary biology*, **8**, 289.

Hoglund J (2009) *Evolutionary Conservation Genetics*. Oxford University Press, New York, NY.

Huang H, Knowles LL (2014) Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic biology*, **0**, 1–9.

IUCN (2012) The IUCN Red List of Threatened Species Version 2012.2.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Kanitz R (2009) Diversidade genética em espécies do gênero Cavia (Rodentia, Mammalia) e a história evolutiva do raro preá de Moleques do Sul. Pontifícia Universidade Católica do Rio Grande do Sul.

Kuhner MK (2006) LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, **22**, 768–770.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lepais O, Weir JT (2014) SimRAD: a R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular ecology resources*, **33**.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–6.

Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lischer HEL, Excoffier L, Heckel G (2014) Ignoring heterozygous sites biases phylogenomic estimates of divergence times: Implications for the evolutionary history of Microtus voles. *Molecular Biology and Evolution*, **31**, 817–831.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 1–14.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, **12**, 443–451.

Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: What can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, **17**, 3428–3447.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, **31**, 1929–1936.

R Core Team (2014) R: A Language and Environment for Statistical Computing.

Rannala B, Yang Z (2003) Using DNA Sequences From Multiple Loci. *Genetics*, 1645–1656.

Reed DH (2010) Albatrosses, eagles and newts, Oh My!: Exceptions to the prevailing paradigm concerning genetic diversity and population viability? *Animal Conservation*, **13**, 448–457.

Ross MG, Russ C, Costello M *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**, R51.

Salvador CH (2006) Biologia da Conservação na teoria e na prática : o estudo de caso de Cavia intermedia , um dos mamíferos mais raros do planeta . Universidade Federal do Rio de Janeiro.

Salvador CH, Fernandez FAS (2008a) Population dynamics and conservation status of the insular cavy Cavia intermedia (Rodentia: Caviidae). *Journal of Mammalogy*, **89**, 721–729.

Salvador CH, Fernandez FAS (2008b) Reproduction and Growth of a Rare , Island-endemic Cavy (Cavia intermedia) from Southern Brazil. *Journal of Mammalogy*, **89**, 909–915.

Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, **46**, 919–925.

Seddon JM, Baverstock PR (1999) Variation on islands: Major histocompatibility complex (Mhc) polymorphism in populations of the Australian bush rat. *Molecular Ecology*, **8**, 2071–2079.

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, **194**, 647–661.

Stolle E, Moritz RF a (2013) RESTseq - Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. *PLoS ONE*, **8**, e63960.

Traill LW, Brook BW, Frankham RR, Bradshaw CJ a. A (2010) Pragmatic population viability targets in a rapidly changing world. *Biological Conservation*, **143**, 28–34.

Xu P, Xu S, Wu X *et al.* (2014) Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *The Plant journal : for cell and molecular biology*, **77**, 430–42.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9264–9269.

Yu X, Sun S (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, **14**, 274.

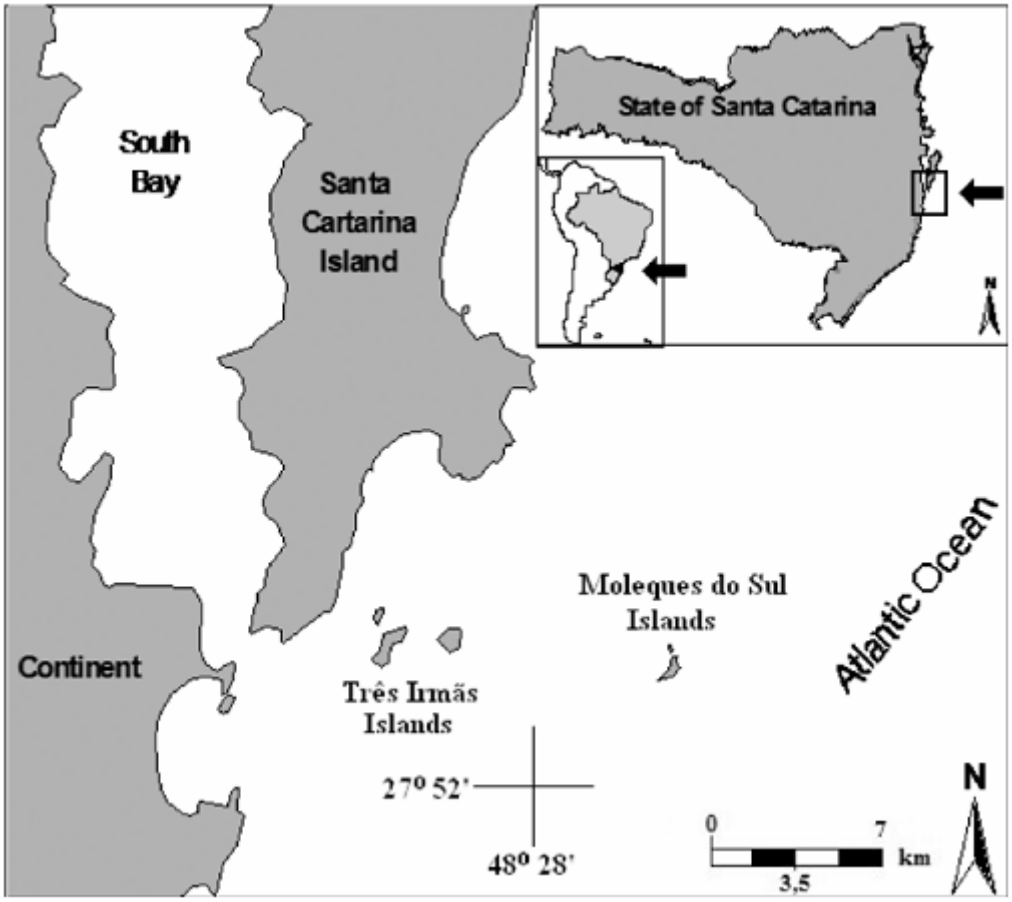

Figure 1: Map of Santa Catarina coast with Moleques do Sul Archipelago (Source: Salvador & Fernandez 2008a)

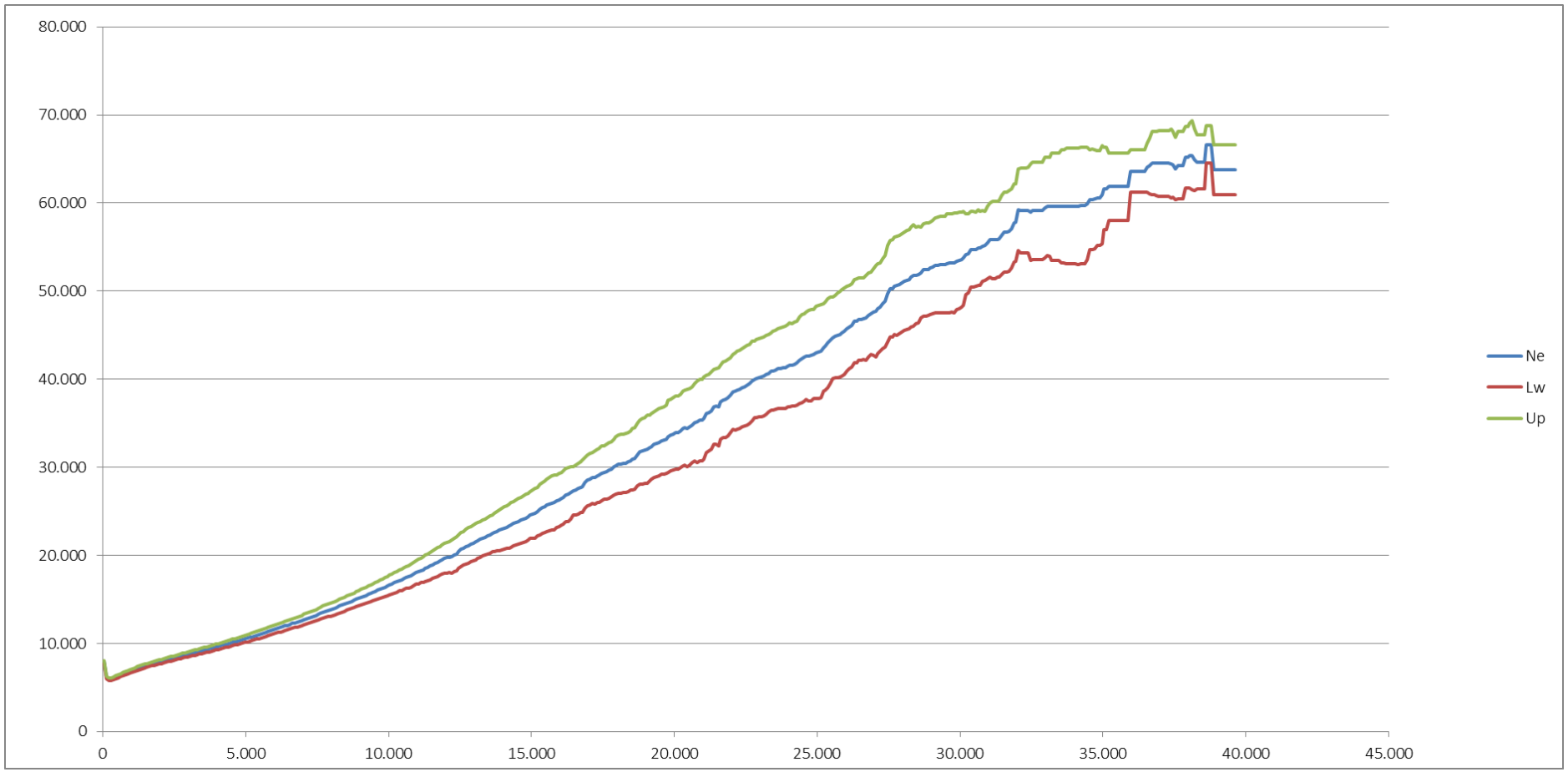

Figure 2: Skyline plot of N_e generated from Migrate-n estimates. Time is scaled in years

	CHIP_1	CHIP_2	CHIP_3	CHIP_4	CHIP_5
TOTAL READS	3.127.968	5.192.400	4.607.304	4.357.964	4.896.432
USABLE (PCT)	60%	64%	72%	60%	63%
MEDIAN	303	267	291	153	274
MODE	326	320	320	295	292

Table 1: Number of analysed loci and observed heterozigosity for each species

	RAD-TAGS		SNP-TAGS	
	C. intermedia	***C. magna***	***C. intermedia***	***C. magna***
# LOCI	71	88	34 446	29 391
# USABLE NUCLEOTIDES	12 396	14 800	20 468	17 400
# POLYMORPHIC	184	205	2 346	4 965
% MISSING DATA	0.2	0.2	0.3	0.3
HETEROZIGOSITY	0.0038	0.00376	0.02377	0.08004

Table 2: Count of usable loci before and after filtering and proportion of genotypes observed. 0/0 and 1/1 represent homozygous loci

	C. INTERMEDIA	*C. MAGNA*	BOTH SPECIES	MEAN
#RAW SNP	811 546	820 949	1 439 196	1 023 897
#PASS FILTER	34 446	29 391	98 569	54 135
MEAN DEPTH	45	45	38	43
MIN DEPTH	7	5	5	6
MAX DEPTH	122	115	232	156
SITES X IND	688 345	587 069	3 940 033	1 738 482
MISSING DATA	205 734	168 707	2 541 189	971 877
0/0	17 760	40 208	214 096	90 688
0/1	8 783	26 642	94 909	43 445
1/1	456 068	351 512	1 089 839	632 473
PROP MISS DAT	29.89%	28.74%	64.50%	41.04%
PROP 0/0	2.58%	6.85%	5.43%	4.95%
PROP 0/1	1.28%	4.54%	2.41%	2.74%
PROP 1/1	66.26%	59.88%	27.66%	51.26%

Table 3: Current and ancestral effectives sizes and estimated divergence time (in years). For theta and HPD values see Supplementary Table S2.

	C. INTERMEDIA	*C. MAGNA*	*ANCESTRAL POPULATION*	DIVERGENCE TIME
BPP	61	90	2 451	55
G-PHOCS	46 719	75 204	25 341 837	143 414
MIGRATE-N	17 245			
LAMARC	11 010			

Supplementary Table S1: Number of sequenced, retained and discarded reads after filtering by *proccess_radtags.pl*. Percentagem of aligned reads is given based on the number of retained reads.

	C_INT.01	C_INT.02	C_INT.03	C_INT.04	C_INT.05	C_INT.06	C_INT.07	C_INT.08	C_INT.09	C_INT.10
TOTAL READS	586 615	369 171	195 344	428 758	628 683	390 052	366 615	617 211	617 512	699 508
RETAINED READS	217 347	161 184	80 433	203 442	209 854	134 653	158 697	295 387	194 614	295 227
DISCARDED READS	369 268	207 987	114 911	225 316	418 829	255 399	207 918	321 824	422 898	404 281
ALIGNED 0 TIMES	47.97%	53.28%	56.84%	55.39%	57.02%	54.19%	52.27%	56.78%	56.71%	55.72%
ALIGNED >1 TIMES	31.88%	35.99%	32.73%	35.70%	33.41%	32.00%	37.18%	33.33%	33.20%	34.36%
ALIGNED 1 TIME	20.15%	10.73%	10.43%	8.90%	9.57%	13.81%	10.54%	9.89%	10.09%	9.92%

	C_INT.11	C_INT.12	C_INT.13	C_INT.14	C_INT.15	C_INT.16	C_INT.17	C_INT.18	C_INT.19	C_INT.20
TOTAL READS	361 221	307 744	433 328	489 552	666 045	436 496	303 847	502 258	627 966	294 341
RETAINED READS	154 342	154 660	118 802	247 887	161 300	172 291	128 871	231 565	301 146	112 267
DISCARDED READS	206 879	153 084	314 526	241 665	504 745	264 205	174 976	270 693	326 820	182 074
ALIGNED 0 TIMES	54.92%	57.52%	51.83%	47.69%	53.10%	55.85%	56.22%	53.57%	55.03%	58.32%
ALIGNED >1 TIMES	34.22%	30.41%	36.47%	30.94%	35.54%	34.65%	34.45%	36.56%	35.12%	32.07%
ALIGNED 1 TIME	10.86%	12.07%	11.71%	21.37%	11.35%	9.50%	9.32%	10.07%	9.85%	9.61%

Cont. Supplementary Table S1

	C_MAG.01	C_MAG.02	C_MAG.03	C_MAG.04	C_MAG.05	C_MAG.06	C_MAG.07	C_MAG.08	C_MAG.09	C_MAG.10
TOTAL READS	406 358	352 715	521 523	610 080	257 898	314 086	319 140	358 934	446 004	375 799
RETAINED READS	127 973	134 972	223 174	245 375	68 898	96 663	110 097	131 358	246 621	184 634
DISCARDED READS	278 385	217 743	298 349	364 705	189 000	217 423	209 043	227 576	199 383	191 165
ALIGNED 0 TIMES	56.90%	58.17%	58.74%	54.88%	52.16%	52.86%	55.66%	56.85%	59.52%	53.20%
ALIGNED >1 TIMES	34.18%	32.78%	32.39%	35.74%	37.82%	37.27%	34.46%	33.95%	32.54%	37.12%
ALIGNED 1 TIME	8.92%	9.05%	8.87%	9.38%	10.02%	9.87%	9.88%	9.20%	7.95%	9.68%

	C_MAG.11	C_MAG.12	C_MAG.13	C_MAG.14	C_MAG.15	C_MAG.16	C_MAG.17	C_MAG.18	C_MAG.19	C_MAG.20
TOTAL READS	199 989	153 034	328 512	434 664	295 168	561 000	807 700	396 496	576 734	415 785
RETAINED READS	91 817	62 389	103 496	151 500	117 135	187 244	316 173	156 884	239 339	119 357
DISCARDED READS	108 172	90 645	225 016	283 164	178 033	373 756	491 527	239 612	337 395	296 428
ALIGNED 0 TIMES	58.83%	56.27%	54.63%	52.90%	57.75%	52.15%	54.10%	54.06%	53.79%	55.75%
ALIGNED >1 TIMES	32.73%	34.35%	35.92%	36.83%	33.54%	37.28%	36.03%	36.04%	36.16%	35.08%
ALIGNED 1 TIME	8.44%	9.28%	9.45%	10.27%	8.71%	10.57%	9.87%	9.90%	10.05%	9.17%

Supplementary Table S2: Theta and Tau estimates for *C. intermedia*, *C. magna* e the Ancestral Population, along with Upper and Lower HPD. N_e is estimates are based on a 4.9 X 10^{-8} mutation rate. Divergence time is given in years, considering the same mutation rate and a generation time of 0.33

	THETA_INT	LHPD	UHPD	THETA_MAG	LHPD	UHPD	THETA_ANC	LHPD	UHPD	TAU	LHPD	UHPD
BPP	1.21E-05	3.00E-06	2.30E-05	1.77E-05	6.00E-06	3.20E-05	4.81E-04	4.18E-04	5.39E-04	8.98E-07	0.0	1.00E-06
G-PHOCS	9.16E-03	1.71E-02	8.35E-02	1.47E-02	3.42E-02	1.38E-01	4.97E+00	4.48E+00	5.47E+00	2.32E-03	8.60E-04	3.81E-03
MIGRATE-N	3.38E-03	3.09E-03	3.67E-03									
LAMARC	2.16E-03	1.75E-03	2.48E-03									

	NE_INT	LHPD	UHPD	NE_MAG	LHPD	UHPD	NE_ANC	LHPD	UHPD	DIV_TIME	LHPD	UHPD
BPP	62	15	117	91	31	163	2 452	2 133	2 750	56	0	62
G-PHOCS	46 719	87 092	425 816	75 204	174 337	704 082	25 341 837	22 877 551	27 882 653	143 414	53 185	235 622
MIGRATE-N	17 245	15 765	18 724									
LAMARC	11 010	8 923	12 658									

Referências Bibliográficas

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular ecology*, **23**, 1661–7.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Computational Biology*, **9**, e1003031.

Cherem JJ, Olimpio J, Ximenez A (1999) Descrição de uma espécie do Gênero Cavia Pallas, 1766 (Mammalia - Caviidae) das Ilhas dos Moleques do Sul, Santa Catarina, Sul do Brasil. *Biotemas*, **12**, 95–117.

Corrêa ICS (1996) Les variations du niveau de la mer durant les derniers 17.500 ans BP : l ' exemple de la plate-forme continentale du Rio Grande do Sul - Brésil. *Marine Geology*, **130**, 163–178.

Dale JW, von Schantz M (2002) *From genes to genomes: Concepts and applications of DNA Technology*. John Wiley & Sons, Ltd, Chichester, UK.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Duran C, Appleby N, Edwards D, Batley J (2009) Molecular Genetic Markers : Discovery , Applications , Data Storage and Visualisation. , **61**, 16–27.

Eaton D a R (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Furnari N (2013) New findings on the origin of Cavia intermedia, one of the world's rarest mammals. *Mammal Review*, **43**, 323–326.

Gava A, Freitas TRO, Olimpio J (1998) A new karyotype for the genus Cavia from a southern island of Brazil (Rodentia - Caviidae). *Genetics and Molecular Biology*, **21**, 77–80.

Hoglund J (2009) *Evolutionary Conservation Genetics*. Oxford University Press, New York, NY.

Huang H, Knowles LL (2014) Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic biology*, **0**, 1–9.

Kanitz R (2009) Diversidade genética em espécies do gênero Cavia (Rodentia, Mammalia) e a história evolutiva do raro preá de Moleques do Sul. Pontifícia Universidade Católica do Rio Grande do Sul.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular phylogenetics and evolution*, **66**, 526–38.

McCormack JE, Maley JM, Hird SM *et al.* (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular phylogenetics and evolution*, **62**, 397–406.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, **12**, 443–451.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Ross MG, Russ C, Costello M *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**, R51.

Rowe HC, Renaut S, Guggisberg a (2011) RAD in the realm of next-generation sequencing technologies. *Molecular ecology*, **20**, 3499–502.

Salvador CH (2006) Biologia da Conservação na teoria e na prática : o estudo de caso de Cavia intermedia , um dos mamíferos mais raros do planeta . Universidade Federal do Rio de Janeiro.

Salvador CH, Fernandez FAS (2008a) Population dynamics and conservation status of the insular cavy Cavia intermedia (Rodentia: Caviidae). *Journal of Mammalogy*, **89**, 721–729.

Salvador CH, Fernandez FAS (2008b) Reproduction and Growth of a Rare , Island-endemic Cavy (Cavia intermedia) from Southern Brazil. *Journal of Mammalogy*, **89**, 909–915.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Stolle E, Moritz RF a (2013) RESTseq - Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. *PLoS ONE*, **8**, e63960.

Yu X, Sun S (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, **14**, 274.