# DE NOVO SNP CALLING AND DEMOGRAPHIC INFERENCE USING TRIO GENOME DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ELİF BOZLAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JULY 2019

Approval of the thesis:

**DE NOVO SNP CALLING AND DEMOGRAPHIC INFERENCE USING TRIO GENOME DATA**

Submitted by ELİF BOZLAK in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics, METU** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU** _____

Asst. Prof. Dr. Aybar Can Acar
Supervisor, **Health Informatics, METU** _____

Assoc. Prof. Dr. Mehmet Somel
Co-Supervisor, **Biological Sciences, METU** _____

**Examining Committee Members:**

Prof. Dr. Tolga Can
Computer Engineering Dept., METU _____

Asst. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Mehmet Somel
Biological Sciences Dept., METU _____

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics Dept., Bilkent University _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU _____

**Date:** **29.07.2019**

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Last name : ELİF BOZLAK**

**Signature : _____**

## ABSTRACT

### DE NOVO SNP CALLING AND DEMOGRAPHIC INFERENCE USING TRIO GENOME DATA

Bozlak, Elif

MSc., Department of Bioinformatics

Supervisor: Asst. Prof. Dr. Aybar Can Acar

Co-Supervisor: Assoc. Prof. Dr. Mehmet Somel

July 2019, 99 pages

De novo mutations are novel mutations which are found in the offspring but not the parents and do not obey the Mendelian inheritance rules. Determining how many de novo mutations occur is important for genetic studies since they help to understand the evolutionary history of populations. In this thesis, we aim to examine de novo mutations that occur within one generation in domestic horses and make estimations on horse demographic history. We used DNA-sequencing data produced by next-generation sequencing technologies from trio data of three different horse breeds: Lipizzaner, Noriker, Haflinger. After quality checks and mapping of the raw data we called genomic variants with three different variant calling algorithms. We filtered all variants depending on their qualities to detect de novo candidates and the final 50 de novo candidates were tested using Sanger resequencing. About 40% of the candidate variants could be validated. We found a higher number of true positives in highly covered Lipizzaner ($n$=13) data, while a lower number of true positives in the low covered Noriker ($n$=3) and Haflinger ($n$=5) data, showing the importance of sequencing coverage to detect true de novo mutations. In addition, we used the Pairwise Sequentially Markovian Coalescent (PSMC) model and performed runs of homozygosity (ROH) analyses to estimate demographic history. Both PSMC and ROH results were coherent with previous studies. All in all, we had an idea for the minimum coverage threshold and quality of whole genome sequencing data, to determine de novo mutations and to estimate population demography.



Keywords: Whole genome sequencing, de novo mutation, Variant calling, Mutation rate, Demography analysis

# ÖZ

## TRIO VERİSİ İLE DE NOVO SNP ÇAĞIRMA VE DEMOGRAFİK GEÇMİŞ ANALİZİ

Bozlak, Elif

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi Aybar Can Acar

Tez Eşyöneticisi: Doç. Dr. Mehmet Somel

Temmuz 2019, 99 sayfa

De novo mutasyonlar ebeveynlerde görülmezken yavruda ortaya çıkan ve Mendel kalıtım kurallarına uymayan mutasyonlardır. Popülasyonların evrimsel tarihlerinin anlaşılmasında yardımcı oldukları için de novo mutasyonaların sayılarının tespit edilmesi genetik çalışmalar için önemlidir. Bu tezde evcil atlarda bir jenerasyonda ortaya çıkan de novo mutasyonları tespit etmeyi ve atların demografik tarihleri üzerine tahminler yapmayı hedefledik. Çalışmada üç farklı at türü (Lipizzaner, Noriker ve Haflinger) için yeni nesil sekanslama teknolojisi ile üretilen üçleme DNA sekans verilerini kullandık. Ham verinin kalite kontrolü ve hizalanmasından sonra, üç farklı varyant çağırma algoritması kullanarak genomik varyantları çağırdık. Tüm varyantları kalitelerine göre filtreledik ve seçilen 50 varyantı Sanger sekanslama ile laboratuarda test ettik. Test edilen varyantların yaklaşık olarak %40'ı valide edildi. Yüksek okuma derinliğine sahip Lipizzaner (n=13) türündeki gerçek pozitif sayısını yüksek, düşük okuma derinliğine sahip Noriker (n=3) ve Haflinger (n=5) türlerindekileri ise daha düşük sayıda bulduk. Sonuçlar gerçek pozitif de novo mutasyonların tespit edilmesinde okuma derinliğinin önemini gösterdi. Ek olarak elimizdeki at popülasyonlarının demografik tarihleri hakkında tahmin yürütmek için PSMC modeli oluşturduk ve ROH analizi yaptık. PSMC ve ROH sonuçları önceki çalışmalarla uyumlu sonuçlar verdi. Sonuç olarak tüm genom sekanslama verisi ile de novo mutasyon tespiti ve popülasyon demografisi tahmini yapabilmek için gereken minimum veri okuması ve kalitesi hakkında fikir sahibi olduk.

Anahtar Sözcükler: Tüm genom sekanslama, de novo mutasyon, Varyasyon çağırma, Mutasyon oranı, Demografi analizi

To My Family

## ACKNOWLEDGMENTS

First of all, I would send my profound thanks to my advisor Asst. Prof. Dr. Aybar Can Acar, my co-advisor Assoc. Prof. Dr. Mehmet Somel and my unofficial co-advisor Dr. Barbara Wallner. They never stop to support me with endless patience and concern during my master's years and they always supported me not just in my academic issues but also in the social issues.

I would also thank my thesis committee members: Prof. Dr. Tolga Can, Assoc. Prof. Dr. Özlen Konu and Assoc. Prof. Dr. Nurcan Tunçbağ, first for accepting to be in my committee and second their precious ideas during the evaluation of my study. In addition to this I owe Tolga Can for the best course that I ever take and his friendly attitudes all the time; I owe Özlen Konu for the positive energy during my thesis period; I owe Nurcan Tunçbağ for her support and trust in many different areas during my master.

I would have special thanks to Dr. Barbara Wallner for providing me the data for my study. She completely trusted and supported me during the analysis of the data. Also, many thanks to Doris Rigler and Eva Michaelis for their time in the laboratory for the validation of the findings.

I would also add my gratitude to Informatics Institute and the 'Music Within' team. Our institute is very supportive and open-minded for new ideas and I saw this attitude at firsthand. My supervisors in 'Music Within' project, Asst. Prof. Dr. Aybar Can Acar and Asst. Prof. Dr. Elif Sürer, never says no to me when I go with the project idea and they always supported me during the project. Besides, Elif Sürer gave me many different perspectives not just for the study and I learned many things from her. A special thanks to Assoc. Prof. Dr. Yeşim Aydın Son and Burcu Üntekin for sharing their data for us for the project. Besides, I would like to thank Yeşim Aydın Son for my great three years in our Bioinformatics Department. I would also thank Ali Çınar for his unrequited help which is an inspiration source to me during the 'Music Within' project. The project was unexpected but very fruitful and fed me in many different aspects during the master.

I also owe special thanks to my family who has an endless support, patience and pure love to me and my ideas all the time: to my brother Barış Bozlak for his pure-minded character, to my mother Nurten Bozlak for her endless support and balancing energy, and my father Metin Bozlak for his strong and promotive personality. There is no doubt that we always try to understand and make the best for each other as shown in my thesis period. Additionally, I believed that they give me a very range perspective on life in many different areas.

I would also send my special thanks to Polat family Suyla Polat, Dilara Polat, Deniz Polat and Haşim Polat for their endless support to me and their endless energy. I can not explain how much I learned from them about life, besides all our music and art sharing. Just luckily, we met again in the Ankara and shared lots of precious moments mostly include brainstormings. Their support during my thesis and also other activities are also nonignorable.

I would have many thanks to my biggest fortunes during my master's years in METU. First, the best colleagues and besties someone can have who are Cansu Demirel, Cansu Dinçer, Evrim Fer, Gökçe Senger, Meriç Kınalı and Muazzez Çelebi Çınar. I cannot tell how I am lucky to know all of them not just in academic cases but also about our sharing on life. Second, my colleagues from the team CompEvo, especially Erinç Yurtman, Reyhan Yaka and Zeliha Gözde Turan. I learned a lot from the team during my master and they supported me all the time. Additionally, I would like to thank my two societies which are METU Capoeira and METU Classical Guitar for many different experiences and the incredible persons they earned to me. We shared many special times and overcome many difficulties with both team and I will never forget the time passed with them. I would send my special thanks to Emre Coşkuner and Burcu Çırtlık for all our special sharing and their endless supports in many different areas during these years. Finally, I would like to thank team MOMO for their promotive attitude and positive vibes during my thesis period.

In the end, I would send special thanks to our department secretary Hakan Güler for his smiling face and patience for us all the time. Because all these people are parts of METU I would send my gratitude to METU and its culture. I can say that, finally I will graduate from a school that I love. I think that METU is one of the precious sides of our country despite it changes with time. I would thank METU for giving me many new perspectives and many precious people.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| **NGS** | Next Generation Sequencing |
| **SNP** | Single Nucleotide Polymorphism |
| **SNV** | Single Nucleotide Variant |
| **PSMC** | Pairwise Sequentially Markovian Coalescent |
| **ROH** | Runs of Homozygosity |
| **TMRCA** | Time to the most recent common ancestor |
| **HMM** | Hidden Markov Model |
| **VCF** | Variant call file |

# CHAPTER 1

# 1.INTRODUCTION

Since horses were domesticated 5000 years ago, they have become a companion to humankind in different areas including transportation, agricultural activities, and wars (Anthony, 2016; Kuderna et al., 2017). Besides these physical advantages, the emotional connection between horses and humans cannot be ignored (Lanata, Guidi, Valenza, Baragli, & Scilingo, 2016). Moreover the domestication of the horses followed by intensive breeding is a very strong example of artificial selection (Zhang et al., 2018), making the evolutionary history of the horse tempting to investigate. Indeed, horses are also chosen as a model organism for the order Perissodactyla (Jagannathan et al., 2018; Wade et al., 2009).

Because of all these reasons above, horses are one of the most well-studied mammals in genetic terms. An updated reference genome of the domestic horse (*Equus caballus*) has recently been released by Kalbfleisch and colleagues (Kalbfleisch et al., 2018). After the release of the updated reference, Jagannathan et al. created a comprehensive variation panel for the modern domestic horse, *Equus caballus* (Jagannathan et al., 2018). In this thesis, we will be analyzing genomic data from modern horses using these resources.

The subject of the thesis will be genetics of three well studied modern horse breeds, namely the Lipizzaner, Noriker and Haflinger. What is common for these horses is that, they all are Austrian breeds. Lipizzaner horses are bred, among other studs, in the Piber stud in Austria and they are selected for dressage performance in the Spanish Riding School (Grilz-Seger et al., 2019). The Noriker is an Austrian draft horse breed and bred for agricultural purposes in the Alps (Druml et al., 2018). And the last breed, which is known as the youngest horse of Austria, is the Haflinger (Druml et al., 2018).

The thesis focuses on the identification of germline ‘de novo’ mutations. ‘De novo’ mutations represent novel mutations occurring in one generation. They are one of the driving sources of gaining new variants throughout generations (Campbell & Eichler, 2013) and they are one of the main contributors of several genetic diseases (Jin et al., 2017). ‘De novo’ mutations are studied from several different aspects in population genetics, as mutation rate per generation is important to study evolutionary history of

populations (Smeds, Qvarnström, & Ellegren, 2016; Tatsumoto et al., 2017). Hence bioinformaticians have spent intense effort on developing 'de novo' mutation detection pipelines and algorithms (Acuna-Hidalgo, Veltman, & Hoischen, 2016; Francioli et al., 2017).

There are two main ways of calculating the 'de novo' mutation rate, either using phylogenetic distance methods or directly detecting 'de novo' mutation candidates in one generation (Campbell & Eichler, 2013). For the well-studied genomes such as that of human, the mutation rate has been calculated via both methods. This provides accuracy in population and medical genetics analyses that depend on the mutation rate. For horses, the only estimate yet has been made by Orlando et al., who calculated the mutation rate for the domestic horse with the phylogenetic distance method. The authors used comparative genomics data from different *Equus caballus* genomes and a donkey genome to estimate the mutation rate (Orlando et al., 2013). Although this rate was used in different horse studies later, there is no mutation rate estimated using 'de novo' mutations for the horse. Providing such an estimate is my main motivation in this thesis.

One application of such rate estimates is in evolutionary demography. Population demography analysis attempts to understand past and present population dynamics, such as bottlenecks, migrations or inbreeding levels (Nielsen & Slatkin, 2013). One way of doing this is using genomic data, and this approach is frequently used to make estimation for the structures of the populations in the past. Pairwise Sequentially Markovian Coalescent (PSMC) is one of these methods to make estimations about population history from genomic data. The algorithm makes estimates of the effective population size at different time points from a single diploid genome (Li & Durbin, 2011). PSMC analysis has been recently used to study horse demographic history as well. Different PSMC studies on horse genomes gave consistent results with each other (Orlando et al., 2013; Schubert et al., 2014). More specifically, these found a decrease in the horse effective population size in the Last Interglacial Period followed by a peak after the Last Glacial Period. The mutation rate used in these studies was again that estimated by Orlando and colleagues (Orlando et al., 2013; Schubert et al., 2014).

Another method used to estimate demographic history is runs of homozygosity (ROH). The results of ROH analysis allow inferences on the amount of inbreeding and presence of bottlenecks in a population in the recent past (Ceballos, Joshi, Clark, Ramsay, & Wilson, 2018; Kirin et al., 2010). Previous works have studied the three horse populations in this study (Austrian Lipizzaner, Noriker and Haflinger) using both pedigree and ROH analyses, and found higher inbreeding in Lipizzaner and Haflinger than in Noriker horses (Druml, Baumung, & Sölkner, 2009; Grilz-Seger et al., 2019, 2018).

In this thesis we analyzed NGS data of trios from the three different domestic horse populations mentioned above. We used data collected from trios, consisting of a mother, a father and an offspring from each of these three breeds. The data were produced by the research group of Dr. Barbara Wallner, Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna. The main aim of this study was to detect 'de

novo' mutations that are found only in the offspring of these breeds. 'De novo' mutations were predicted and mutation rates per generation were inferred for the domestic horse. This study is thus the first to calculate horse mutation rate per generation, inferred from genome-wide 'de novo' mutations. We further used the data to make demographic estimations for demography of these three populations using PSMC and ROH analyses, and the results were compared with those of the previous studies.

The second chapter of the thesis describes the importance of 'de novo' mutations in the fields of genetics and bioinformatics, as well as giving examples of mutation rate estimates and their calculation methodologies. Besides, we compare the different bioinformatics algorithms that we used for the analyses, especially for variant calling. Finally, we mention working strategies of PSMC and ROH methods and findings from different studies for domestic horse populations.

The third chapter explains our methodology to analyze whole-genome NGS data in detail to detect 'de novo' mutations produced in one generation. Here we used different calling algorithms and two different reference genomes. The chapter also details how we performed PSMC and ROH analysis.

The fourth and fifth chapters describe our results and compare them with previous studies. Then, we make inferences on stability of our results and discuss possible limitations and future directions.

# CHAPTER 2

# 2.LITERATURE REVIEW

## 2.1. Mutation rates and 'de novo' mutations in evolutionary studies

Germline mutations are one of the fundamental forces of evolutionary mechanisms due to the their contribution to the variation pool of populations (Loewe & Hill, 2010). Besides, they are also one of the causes of heritable and complex diseases (Shendure & Akey, 2015). These properties necessitate making accurate predictions on mutation rates (Narasimhan et al., 2017; Shendure & Akey, 2015). In addition to some recent population genetic methods, there are two main ways of estimating mutation rates. The first one is calculating mutation rates based on phylogenic distances. In this method, comparative genomic data from two lineages with known splitting time (based on the fossil record) is used to calculate the nucleotide divergence between the two. This divergence can then be used to estimate the per generation substitution rate, which can also be used as the per generation mutation rate. This approach is based on the neutral theory of molecular evolution, which states that for neutral loci, the substitution rate should be equal to the mutation rate (Nielsen & Slatkin, 2013). Furthermore, it assumes that the generation time is known and fixed for the lineages in question. The other method for calculating the mutation rate is using 'de novo' mutations. In this approach, direct calculations from genomic data of a trio (including mother, father and offspring data) is used and the number of mutations only observed in the offspring is divided by the genome length (Campbell & Eichler, 2013; Narasimhan et al., 2017; Shendure & Akey, 2015).

There are several studies for mammals which have used the phylogenetic distance method for mutation rate calculations. For instance, the mutation rate was estimated as ~2e-08 per base per generation for human and nonhuman primates by considering divergence between neutral sites in their genomes following the above approach (Mikkelsen et al., 2005; Shendure & Akey, 2015). The rate has been reestimated as 1e-09 per base per year with the same method in different studies (Prüfer et al., 2012) by assuming the divergence time between human and chimpanzee as 6-7 million years ago (Ma). For the domestic

horse, Orlando et al. estimated the mutation rate per generation with the phylogenetic distance method (Orlando et al., 2013). They estimated the mutation rate as 7.2e-09 per site per generation by comparing horse and donkey, and assuming that donkey and horse diverged 4-4.5 Ma (Janečka et al., 2018; Jónsson et al., 2014). Both primate and equid phylogenetic mutation rate estimates thus appear roughly within the same order of magnitude (the estimate for equids being midway between different estimates for primates).

Although the phylogenetic distance method has been frequently used in many different studies to estimate the mutation rate, scientists have pointed out multiple weaknesses of this method. Tatsumoto et al. suggested that different factors such as generation time, effective population size, rate of heterogeneity causes uncertainty in results when using the phylogenetic approach (Tatsumoto et al., 2017). Francioli and his team also argued for the disadvantages of the method because different selection mechanisms are effective on different populations and species, and across different regions of the genome, and therefore identifying truly neutral regions will not be straightforward (Francioli et al., 2014).

Given these problems of the mutation rate estimation based on phylogenetic distance, and thanks to the availability of high throughput (next generation) sequencing technologies, scientists in recent years have started making direct estimates of the mutation rate by ‘de novo’ mutations scans. These studies use high quality whole genome data from trios (for sexually reproducing species) produced by next generation sequencing. To date, mutation rate estimations from 'de novo' mutations have been made in a few mammalian species, such as humans, chimpanzee and mouse. While some studies compared different families of the same population (Francioli et al., 2014; Jónsson et al., 2017), others focused on families from different populations (Conrad et al., 2011). For example, Jonsson and colleagues worked with a large genomic data dataset compromising 1,548 human trios from Iceland. They made estimations on the 'de novo' mutation rate in one generation and they estimated the rate as 1.28e-08 per base per generation. Meanwhile, Francioli and colleagues analyzed genomic data from 250 Dutch parent-offspring families including trios and twin-parent families, and they calculated mutation rates for different parts of the human genome to understand possible explanations for the different mutation rates. Conrad and colleagues calculated and compared mutation rates inferred from trios of European descent ($u$=1.17e-08 bp/gen) and Yoruban descent ($u$=0.97e-08 bp/gen) and found possible sources for the differences between two populations. In addition to human studies, nowadays 'de novo' mutation studies are also being conducted in other mammalian species. Tatsumoto et al. produced ultra-deep sequencing data for a chimpanzee trio to calculate 'de novo' mutation rate for the chimpanzees, and they estimated this as 1.48e-08 per site per generation (Tatsumoto et al., 2017). Another popular mammal subjecting for mutation rate calculations is mice (*Mus musculus*). Uchimura and colleagues calculated the germline mutation rate for mice is 5.4e-09 bp/gen (Uchimura et al., 2015) and Lindsay and colleagues calculated the rate as 3.9e-09 (Lindsay, Rahbari, Kaplanis, Keane, & Hurles, 2016). Both rates were calculated using the number of ‘de novo’ mutations detected in the offspring but not in the mouse population of interest. Intriguingly, the

mutation rate of humans appeared almost 2 to 3 times higher than the mouse. This difference has been explained, by Lindsay and colleagues, by the larger effective population size of the mouse (and thus stronger selection for high fidelity DNA copying) when compared with the human (Lindsay et al., 2016). Notably, the resulting rates from these studies were comparable with each other irrespective of populations and species, and also consistent with the rates that were calculated with the phylogenetic distance methods for human, all being on the order of ~1e-08 bp/gen.

Narasimhan and colleagues have discussed possible reasons for the discrepancies between mutation rates estimated with the two different methods (phylogenetic distance vs. direct estimations from 'de novo' mutations). They suggested different explanations such as i) sample size (including number of individuals used in the study or number of true 'de novo' mutations); ii) using only mutations seen in one generation; iii) effect of paternal age; iv) genomes of diseased individuals (Narasimhan et al., 2017). Additionally, Conrad et al. showed that, although mutation rates were calculated with the same methods in different trios of same species, differences in these rates could also be observed (Conrad et al., 2011). They suggested different reasons for this difference. The first one is variation due to differences in genetic background and selection dynamics among individual. For instance, although it is accepted that more mutations are occurring in the paternal lineages (Haldane, 1947), Conrad et al. found a different pattern in one of their studied families. In this family a higher percentage of 'de novo' mutations derived from the maternal lineage. The second possible reason is the age of parents. This idea is also supported by different studies. Kong and team showed that the increasing age of the father results in an increase in the number of 'de novo' mutations occurring in one generation (Kong et al., 2012), a conclusion also supported by Francioli et al. (Francioli et al., 2014). In addition to these reasons, Francioli et al. found that different genomic regions could have different mutations rates. This can create a bias in mutation rate estimates from whole genome data depending on the distribution of number of reads mapped along the genome.

## 2.2. Importance of data amount and quality in next-generation sequencing analysis

### *2.2.1. Quality of whole-genome data*

Different studies show the importance of sequencing depth in detecting true 'de novo' mutations. Tatsumoto and team estimated mutation rate with their ultra-deep high-quality sequencing data (Tatsumoto et al., 2017). To do this, they validated their candidate 'de novo' SNPs by Sanger sequencing. They categorized their data as 30x, 60x, 90x and 120x groups, dependent on the sequencing depth. Then they took the 90x and 120x candidate lists and found all true positive mutations ($n$=32) in the shared list of the two categories. On the other hand, there were 9 candidates found only in 90x data, and these were all detected as false positives in Sanger analysis. Based on these results, they suggested that even 90x coverage is not enough to eliminate all false positives. According to this study, determining true positive 'de novo' SNPs, which includes eliminating false positives and

avoiding false negatives, can only be revealed with ultra-deep covered data (Tatsumoto et al., 2017). 'De novo' mutation studies in human trios have also supported these results. Jonsson et al. used data with an average of 35x coverage and Conrad et al. used data greater than 22x coverage, and their mutation rate estimates per generation (1.28e-08 in Jonsson et al. and ~1.2e-08 in Conrad et al.) were consistent with each other (Conrad et al., 2011; Jónsson et al., 2017). Although, Tatsumoto et al. suggested that these coverages are not sufficient to detect 'de novo' mutations, Jonsson detected on average 70.3 'de novo' mutations in 1548 families trio data and Conrad detected 49 and 35 'de novo' mutations in the two trios they worked with. This is an indication of uncertainty about the lower data thresholds to detect 'de novo' mutations.

### *2.2.2. Quality of reference genome*

The quality of the reference genome is another factor that affects mapping rates, variant calls and consequently 'de novo' mutation rate estimates (Kalbfleisch et al., 2018; Li & Wren, 2014). An increase in the mapping rates was seen in the updated human reference genome from GRCh38 when compared with GRCh37 (Guo et al., 2017). Besides this increase in the mapping, Guo et al. also detected fewer number of SNVs in the GRCh38. They explained the reason for this lower number of variants using GRCh38 by fewer false positive variants due to quality of new reference. In addition, Li and Wren compared human references GRCh37 and GRCh38 in their analysis and showed the effects of different parameters in variant calling (Li & Wren, 2014). They also found fewer false positive variants when using the updated reference.

In the case of horse, the first high-quality reference assembly for the domestic horse, EquCab2, was released by Wade et al. (Wade et al., 2009). The updated version of the *Equus caballus* reference genome was recently published in 2018. The data which were used to construct the updated reference was derived from the same Biosample, the female Thoroughbred 'Twilight'. In the updated version of the reference genome EquCab3 different sequencing and assembly technologies were used. In the updated reference genome, the number of non-N bases on the chromosomes were increased to 2.41 Giga base (Gb) from 2.33 Gb in EquCab2. Besides, the gaps in the incorporated chromosomes decreased 10-fold in the EquCab3. To test the updated reference, whole genome sequencing data from two Thoroughbred male horses were mapped to both EquCab2 and EquCab3 references (Kalbfleisch et al., 2018). These authors found an increased number of mapped reads and properly paired mapped reads on the EquCab3 mapped data when compared to that of EquCab2. Kalbfleisch et al. suggested that this might be a result of the improvement in the updated genome in several ways, such as containing fewer gaps (Kalbfleisch et al., 2018). All in all, these comparative studies suggest that an updated reference genome that includes deep coverage high throughput sequencing data, may provide more accurate results.

### *2.2.3. Quality and properties of existing variation panels*

The importance of variation panels in the 'de novo' mutation detection workflows was previously mentioned by Romero and colleagues (Gómez-Romero et al., 2018). These authors suggested that, because there is a low probability for the occurrence of independent 'de novo' mutations at the same position in different individuals, standing variation can be used to eliminate false positives, which leads then to more accurate 'de novo' mutation estimates. They filtered their candidates in the study, based on this idea. In the recent past, 50k, 70k and 670k SNV variation panels were produced for targeted genotyping arrays with previous reference genomes of the domestic horse (McCoy & McCue, 2014; McCue et al., 2012; Schaefer et al., 2017). A variation panel produced from whole genome sequencing data of 88 horses from different breeds was recently released (Jagannathan et al., 2018). In this panel the updated version of the reference (EquCab3) was used as reference genome. Jagannathan et al. called approximately 23.5 million SNVs in this study. This newly constructed horse variation panel based on EquCab3 by Jagannathan et al. can also be used as an efficient tool to eliminate false positives to detect true 'de novo' candidates in horse trios.

## 2.3.Analyzing high throughput sequencing data for 'de novo' mutation detection

### *2.3.1. Differences among variant calling algorithms*

There are several tools developed to find variations from the reference genome and variations among different individuals. In recent years multiple studies have compared the performances of different variant callers. Poplin et al. has recently reviewed different variant calling algorithms and compared their performances (Poplin et al., 2017). These algorithms can be split into two main groups. One group finds mismatches between the subject individual and the reference genome. These are called 'pileup' callers, and SAMtools (Li et al., 2009) and GATK's UnifiedGenotyper (UG) (Depristo et al., 2011) are examples of this group. Despite their high sensitivity, these algorithms have some weak points caused by using a reference sequence to find variants (Rimmer et al., 2014) or considering one position at a time during variant calling process (Garrison & Marth, 2012; Rimmer et al., 2014). The other group of variant caller algorithms are called assembly-based algorithms; these basically construct haplotypes by creating de Bruijn-like graphs. Platypus (Rimmer et al., 2014), GATK's HaplotypeCaller (HC) (Poplin et al., 2017) and FreeBayes (Garrison & Marth, 2012) are examples of this group of variant callers (Poplin et al., 2017; Xu, 2018). These tools first assemble reads locally and produce candidate haplotypes; then they estimate likelihood of haplotypes by aligning the reads to the haplotypes one by one and counting them. The advantages of these latter algorithms becomes most obvious in highly variable genomic regions, because they do not depend on local alignment as 'pileup callers' do, which causes mistakes in this high variable regions (Li & Wren, 2014; Poplin et al., 2017; Rimmer et al., 2014). In addition, assembly-based algorithms can detect the co-existence of different variants at the same time (Xu, 2018). Poplin and team compared variant callers including both types of algorithms

('pileup' and assembly-based) and among four algorithms, namely GATK's HaplotypeCaller (HC), GATK's UnifiedGenotyper (UG), Platypus, SAMtools, they found the highest sensitivity in HC. The authors also suggested that assembly-based algorithms give more accurate results. Another idea that is also proposed in the HaplotypeCaller's paper is that calling variants jointly in multiple individuals improves accuracy of the calling results, especially for the regions with low coverage (Poplin et al., 2017; Rimmer et al., 2014), but such joint calling results in the production of more complex de Bruijn graphs, so it needs more computational power (Poplin et al., 2017). HaplotypeCaller, Platypus and FreeBayes try to solve this computational complexity in the algorithms while they make haplotype-based variant calling.

Sandmann and colleagues used and compared GATK's HC, Platypus and FreeBayes besides other variant callers on their targeted Illumina HiSeq data (Sandmann et al., 2017). The main aim of their study was to produce a fast pipeline to use in clinical applications to identify both rare and common variants. After analyzing their data with different variant calling algorithms, they reported that Freebayes is the most sensitive among these three variant callers, despite reporting many false positive variants. They further reported that the accuracy of Platypus's results depends on allele frequency of the variant and that of HaplotypeCaller depends on the type of platform, specifically whether HiSeq or NextSeq was used to produce the data (Sandmann et al., 2017).

Some algorithms like PhaseByTransmission (PBT) (Francioli et al., 2017) are designed to directly use known pedigree information of trio data and possibly allele frequency in the population, which can be further used to detect 'de novo' mutations. PBT calculates posterior probabilities of each candidate variant and outputs a list of candidates (Francioli et al., 2017). Francioli and team compared their PBT algorithm with other 'de novo' detection tools, namely DeNovoGear (Ramu et al., 2013) and TrioDeNovo (Wei et al., 2015). They found consistent results using high coverage data, but they found PBT results to be more accurate using relatively low coverage data downsampled to 15x. On the other hand, PBT found many false positive 'de novo' candidates which could not be validated in the low coverage data (2-20x) in all trio members.

### *2.3.2. Importance of filtering variants*

Filtering variant lists produced by calling algorithms is a mandatory step to eliminate false positive variants as much as possible. Li and Wren made a comprehensive analysis to show how the filtering steps affects the quality of analysis (Li & Wren, 2014). They used FreeBayes, GATK UnifiedGenotyper, Platypus and GATK HaplotypeCaller for variant calling in their analysis and they filtered their data depending on different parameters: low-complexity regions, maximum depth, allele balance, double strand filter, Fisher strand filter, quality filter. The authors found that although the low complexity region filter is the most effective filter to eliminate false positive heterozygous variants overall, when using data aligned with the BWA-mem, the maximum depth filter is the most effective filter among other filters. The other parameters gave different result for the different call sets (Li & Wren, 2014). One of the other well-accepted filtering strategies is GATK's hard

filtering strategy which is suggested by Summa and colleagues (De Summa et al., 2017). These authors tried to identify parameters for 7 standard GATK filters followed by a classification tree to decide correct variants.

Besides the general filters mentioned above, there are several other strategies to detect 'de novo' mutations using genomic variation data. Jonsson et al. used minimum depth, allele balance and genotype likelihood parameters in their works to define their 'de novo' mutation candidates (Jónsson et al., 2017). Tatsumoto et al. also used read depth and allele balance in their 'de novo' SNV defining strategy (Tatsumoto et al., 2017). Li and Wren mentioned the importance of depth and allele balance in filtering (Li & Wren, 2014). While positions having low depth (depending on the study) are not significant, positions with high depth are not efficient for the detection of true 'de novos' (Jónsson et al., 2017; Li & Wren, 2014; Tatsumoto et al., 2017). Allele balance filter is also meaningful to eliminate mutations which show low proportions of alternative allele to reference allele counts, which is indicative of technical errors (Li & Wren, 2014).

## *2.4.* Demography analysis in *Equus caballus*

Different estimations on demography such as population size, migration, inbreeding or history of divergence could be made with population genetics data (Nielsen & Slatkin, 2013). With the increase in the amount and abundance of genomic data, new methods are being developed to make estimations on the demographic history of populations. Estimating evolutionary trees or making cluster analysis between individuals of a population could be some examples for demography analysis. Here we used two population genetics methods, pairwise sequentially Markovian coalescent (PSMC) and runs of homozygosity (ROH), to study horse demographic history.

### *2.4.1. Pairwise Sequentially Markovian Coalescent (PSMC)*

PSMC is a haplotype-based method to make estimations on population demography from genomic information (Chen, 2015; Li & Durbin, 2011). The main idea behind the method is that, using whole genome diploid sequence data from a single individual, changes in past effective population sizes can be estimated. It uses the proportion of heterozygote sites in a genomic segment separated from other segments by ancestral recombination events and infers the local time to the most recent common ancestor (TMRCA) by using a Hidden Markov Model (HMM). In this HMM model the observation is the diploid sequence of an individual and the hidden states are the TMRCA and ancestral recombination events, the latter represented by transitions. The parameters that are necessary for the calculations are mutation rate, recombination rate and generation time in the population, which are estimated from the data using an expectation-maximization algorithm. Results of the analysis of human whole genome data provides estimates of the changes in the population size between 20 kya to 3 Mya. Getting information for a larger time period is also possible but Li and Durbin suggested that it will not be giving accurate results (Li & Durbin, 2011).

There are different studies which investigated horse demography by the PSMC method. Orlando and colleagues studied genomes of an ancient horse, five modern domestic horses, a Przewalski's horse and a donkey with PSMC analysis. First, they calculated a mutation rate for the horse genome as 7.2e-09 per site per generation by aligning genome sequences of the horse and the donkey. They assumed 8 years generation time for the horses. Then they performed PSMC analysis using the calculated mutation rate and generation time. They made estimations for the last 2My of the horse population (Orlando et al., 2013). Their results suggested that the horse population size reached a minimum around 125 kya before present, a time period that coincides with the Last Interglacial Period. On the other hand, the population size reached the maximum level between 25-50 kya, which corresponds to climatic changes fruitful for the vegetation after the Last Glacial Maximum (Lorenzen et al., 2011). Schubert and team also made PSMC analysis for the horses by using parameters from Orlando et al.'s 2013 study. They analyzed another two ancient horse genomes, again together with domestic and Przewalski's horse genomes, to detect signals of domestication by using whole genomes of the horses and found patterns similar to the findings of Orlando and colleagues' study (Schubert et al., 2014).

Although the PSMC analysis gives highly accurate estimations, Orlando et al. also detected a bias in the result for low coverage data (<20x) (Orlando et al., 2013). Nadachowska-Brzyska and colleagues also mentioned the impact of data coverage in the PSMC analysis. They analyzed 200 individuals from four different flycatcher species and made PSMC analysis to have idea of demographic history of their populations. They found that their results are much more consistent using data from individuals having at least 18x mean coverage and applying a 10x coverage filter per-site. They suggested to perform PSMC analysis using this approach (Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016).

### *2.4.2. Runs of homozygosity (ROH)*

Calculating runs of homozygosity is another technique to detect bottlenecks and inbreeding rate in populations. The idea underlying this technique comes from the distribution of autozygous sites in genomes. Broman and Weber proposed the idea that these sites are not evenly distributed in the genome, but they are dispersed in the form of runs, or tracks (Broman & Weber, 1999; McQuillan et al., 2008). Cebellos and team recently discussed the relationship between ROH sites and population dynamics in a review (Ceballos et al., 2018). First, they showed that the length and distribution of ROHs in a genome are related with population size. The relationship occurs such that smaller populations have longer and more ROHs than larger populations. Then the authors explained how different processes affect the distribution of ROHs. Admixture events cause decrease in the length and number of ROHs, whereas bottlenecks lead to an increase in the number of ROHs but usually keep the length relatively short. Inbreeding is shown to result in an increase in the number of very long ROHs. If a population both went through a bottleneck and exercised inbreeding, the individuals will show both many short ROHs and some very long ROHs (Ceballos et al., 2018).

The horse (*Equus caballus*) is well studied for ROHs by many different groups. Grilz-Seger et al. made ROH analysis with four different sub-populations of Lipizzaner horses (Grilz-Seger et al., 2019). One of the conclusions the authors reached is Austrian Lipizzaner being more inbred than other Lipizzaner populations. They showed that cumulative ROHs ($F_{ROH}$) in different lengths was almost twice in the Austrian Lipizzaners compared to Croatian, Hungarian, and Slovakian Lipizzaners. In addition, the number of long ROHs (>4Mb) were two to ten times higher in Austrian compared to the other Lipizzaners. Based on these results they suggested Austrian Lipizzaners are more inbred and bottlenecked (Grilz-Seger et al., 2019). Furthermore, they estimated inbreeding coefficients from ROH data, and these estimates were also consistent with the inbreeding coefficients calculated by pedigree analysis for the Austrian Lipizzaner (Zechner et al., 2002). Grilz-Seger therefore suggested that the Austrian Lipizzaner is the most inbred sub-population among Lipizzaners. In addition to the ROH analysis, Druml et al. analyzed pedigree data of Austrian Noriker horses and they suggested that inbreeding in Austrian Noriker is lower than Austrian Lipizzaner (Druml et al., 2009). However, the inbreeding coefficient calculated for the Austrian Haflinger, which was calculated by Druml and team in 2018, was much higher and similar to the one observed in the Austrian Lipizzaner (Druml et al., 2018; Grilz-Seger et al., 2019).

# CHAPTER 3

# 3.MATERIALS AND METHODS

Here, we explain the methodologies using in this study, which covers whole genome sequencing data analysis including read mapping and variant calling, as well as demography analysis.

## 3.1. Next Generation Sequencing Data

Whole-genome Illumina Next Generation Sequencing (NGS) data for three family trios (in total nine horses) from different breeds were provided by Dr. Barbara Wallner, Institute of Animal Breeding and Genetics, University of Veterinary Medicine in Vienna, Austria. All trios include mother, father and a male offspring. Trio 1 were three purebred Lipizzan horses, trio 2 purebred Noriker horses and trio 3 purebred Haflinger horses. Whole blood samples of each individual were collected for NGS data generation in 2012.

NGS data was generated in different years and on different platforms. The Lipizzaner trio was sequenced by HiSeq2000 at BGI (Beijing Genomics Institute) in 2013, whereas the Haflinger and Noriker data were produced by HiSeqV4 at the Core Facility for Sequencing at the Vienna BioCenter (CSF) in 2015. The data consist of paired-end reads. Read length in Lipizzaner reads was 90 base pairs (bp), whereas the read length in Noriker and Haflinger reads was 125 bp. General information about the data is given in Table 3.1 below.

Haflinger and Noriker data were provided as unmapped BAM files, while Lipizzaner data in FASTQ format. Before starting the quality analysis of raw files, all BAM files were sorted with SAMtools (Li et al., 2009) (see Appendix A for version and command line usage of the tool) and converted to paired-end FASTQ files by BEDTools' (Quinlan & Hall, 2010) (see Appendix A) ‘bamtofastq’ algorithm.

**Table 3.1:** General information of individuals and provided data formats of each individual. Some of the individuals were sequenced on multiple lanes.

| Sample ID | Breed | Pedigree | Number of sequencing lanes | Initial file format |
|---|---|---|---|---|
| **111** | Lipizzaner | Son | 3 | FASTQ |
| **113** | Lipizzaner | Father | 3 | FASTQ |
| **166** | Lipizzaner | Mother | 4 | FASTQ |
| **BW-352** | Noriker | Father | 2 | BAM |
| **BW-353** | Noriker | Mother | 1 | BAM |
| **BW-354** | Noriker | Son | 1 | BAM |
| **BW-355** | Haflinger | Father | 1 | BAM |
| **BW-356** | Haflinger | Mother | 2 | BAM |
| **BW-357** | Haflinger | Son | 1 | BAM |

## 3.2. First Data Analysis

### *3.2.1. Quality control and trimming of FASTQ files*

Quality controls of the FASTQ files from each sample were performed with FastQC (Andrews, 2015) (see Appendix A). Then, all reads were quality-trimmed with Trimmomatic (Bolger, Lohse, & Usadel, 2014) (see Appendix A) using the following parameters: "TRAILING:10, MINLEN:50, SLIDINGWINDOW:5:20". These parameters were chosen due to the low-quality reads, some of which were under Phred score of 30 towards the end of the reads. Next, the quality of all trimmed FASTQ files was reinvestigated using FastQC. To visualize FastQC results, the MultiQC software (Ewels, Magnusson, Lundin, & Käller, 2016) (see Appendix A), which creates a report by summarizing FastQC results, was used.

### *3.2.2. Mapping reads to the reference genomes*

First, reference genomes EquCab2 (Wade et al, 2009) and EquCab3 (Kalbfleish et al., 2018) were indexed with BWA (Li & Durbin, 2010) using the 'bwtsw' algorithm. Then the 'mem' (Li, 2013) algorithm of the BWA, which shows high performance in mapping of long reads was used for mapping reads to the reference genomes. As for some

individuals the FASTQ data was derived from different sequencing lanes (see Table 3.1), data from each lane were mapped to the references separately. SAM files were obtained at the end of the mapping, which were converted to binary BAM files with the SAMtools' 'view' algorithm and these BAM files were sorted with the SAMtools' 'sort' algorithm. For individuals with more than one lane sequenced, each lane was treated in the same way and then data from different lanes were merged with the SAMtools' 'merge' algorithm to create a merged BAM file, which contains data from separate lanes. These BAM files were sorted again, and only properly paired reads were chosen with the SAMtools' 'view' algorithm using the parameters: "-F 4 -f 2 –h". Duplicate reads were removed by the MarkDuplicates algorithm of Picard Tools with the "REMOVE_DUPLICATES=TRUE" parameter. At the last step, reads that were under quality score 20 were filtered out with SAMtools' 'view' again, with the '–q 20' parameter. Insert sizes of the produced BAM files that were mapped to the EquCab3 were calculated by Picard Tools's CollectInsertSizeMetrics function. Finalized BAM files were indexed with SAMtools. Versions of all tools and their command line usages were given in Appendix Table A.A.1 and A.A.2.

### *3.2.3. Calculating mapping coverage*

Coverages of the mapped BAM files were calculated with the 'genomecov' function of BEDtools (see Appendix A) for each horse. Then, the mean mapping on autosomal chromosomes for each individual and also for each trio were calculated in R (see Table A.A.1 for version).

To identify positions within a certain depth range, positional coverages of each individual were also calculated using the same function of BEDtools with the "–d" parameter (see Appendix A). The resulting coverage files were separated into chromosomes by UNIX command-line functions. Then, an R code that checked whether a certain position was within the given depth range in each member of the respective trio was written. The thresholds were set to 10 to 30 reads for EquCab2 mapped data, and 10 to two times mean coverage of the respective trio in the EquCab3 mapped data (Li & Wren, 2014).

## 3.3. Variant calling

Three different algorithms were used to call variants including single nucleotide variants (SNV) and insertion-deletion variants (INDEL) in both references. These were i) GATK's 'HaplotypeCaller' (Poplin et al., 2017) function, ii) Platypus' (Rimmer et al., 2014) 'callVariants' function, and iii) the Freebayes (Garrison & Marth, 2012) algorithm (see Appendix A for version and command line usage of the tool). Default parameters of each callers were used, and members of the same trio were given to the callers at once while performing variant calling. The output variant calling file (VCF) of each algorithm included all variants detected in the trio horses when compared to the reference genome. At the end of variant calling step, six different VCF files had been generated per trio, one from GATK, one from Platypus and one from Freebayes, for each reference.

### *3.3.1. Workflow for determining candidate 'de novo' mutations*

In the next step, different filters were applied to detect 'de novo' mutations, which we define as occurring in the offspring and not found in the parents. To predict 'de novo' mutation candidates, a custom Python (see Appendix Table A.A.1 for the version) code was written. The code parsed each variant in the VCF file line by line. First, it checked the variant for any undetermined genotype among the three individuals. Such variants were filtered out. In next step, the code compared the offspring's genotype with its parents' genotype. Table 3.2 shows all of the conditions that were chosen as a 'de novo' candidate in the offspring. The first four rows of the table show the homozygous cases and last two rows show the heterozygous cases in the offspring.

**Table 3.2:** 'de novo' candidate patterns extracted from VCF files.

| Offspring | Parent 1 | Parent 2 |
|---|---|---|
| 0/0 | 1/0 | 1/1 |
| 0/0 | 1/1 | 1/0 |
| 1/1 | 0/0 | 1/0 |
| 1/1 | 1/0 | 0/0 |
| 0/1 | 0/0 | 0/0 |
| 0/1 | 1/1 | 1/1 |

After choosing 'de novo' candidates, chrX, chrUn, chrM, and INDELs (insertions and deletions) were removed by another Python code. At the end, the output files contained SNV 'de novo' candidates which are located on the autosomal chromosomes. For the final 'de novo' list for each trio, 'de novo' SNP candidates called by three different variant calling algorithms were considered. For this we generated BED files from the VCF files using Bedtools (see Appendix A) and then the 'intersect' function of Bedtools to choose shared variants. These final position files were converted into VCF files with a Python script. While creating VCF files from BED files, genotype information for VCF files was derived from the GATK variant file of each related trio.

For the next filtering steps, the GATK tool was used. First, the allele balance of each variant was calculated by the 'VariantAnnotator' function, with the '-A AlleleBalanceBySample' parameter. Then GATK hard filtering (De Summa et al., 2017) was applied to the VCF data using 'VariantFiltration' function with the given parameters: "QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, QUAL < 30, AC > 4". 'de novo' candidates that passed filtering were kept and heterozygous candidates in each offspring were selected among these with the '.isHet()'

function. The last two columns of Table 3.2 show the possible allele distribution of selected variants among trio individuals. In the last step, the genotype statistics (including GT, AD, DP, GQ and PL) of individuals in each trio were filtered with 'VariantFiltration' and '--genotypeFilterExpression' functions. Filtering parameters for genotypes were: AB < 0.3 || AB > 0.75, GQ < 40.0 (Jónsson et al., 2017; Maretty et al., 2017). For the depth (DP) filtering, we required depth between 10x and 30x for the EquCab2 mapped data. In EquCab3 mapped data, we changed the threshold and we required lower and upper thresholds between 10x and 2 times of breed mean coverage, respectively (Li & Wren, 2014). Analysis steps are shown in Figure 3.1 as four main parts (Mapping, Variant Calling, Filtering, Comparison). Versions of tools and their command line usages are given in Appendix Table A.A.1 and A.A.2.

### *3.3.2. Determining candidate 'de novo' mutations with GATK PhaseByTransmission (PBT) algorithm*

To test the results of the workflow in section 3.3.1, GATK's PhaseByTransmission (Francioli et al., 2017) algorithm, which predicts candidate 'de novo' mutations, was implemented as an alternative strategy. The function was run with the Lipizzaner variants from three different algorithms (GATK HaplotypeCaller, Platypus and Freebayes) separately. The '--MendelianViolationsFile' parameter was added to obtain candidate 'de novo' mutations as a list. Shared 'de novo' candidates from the three different calling algorithms were selected. GATK's PhaseByTransmission (see Appendix A) algorithm was applied only for the Lipizzan trio and in the next step (3.3.3) the resulting list of 'de novo' candidates was compared to the ones' predicted with the custom approach described in 3.3.1 (workflow for determining candidate 'de novo' mutations).

### *3.3.3. Comparing different variant lists*

In the next step, 'de novo' candidates determined by the workflow described in subsection 3.3.1 for both references were compared with each other and with Jagannathan and colleagues' (Jagannathan et al., 2018) domestic horse variant list including 23.5 million SNVs. Candidates that were shared between lists and candidates unique to only one list were determined. At the same time, 'de novo' candidates produced by the PBT algorithm (see subsection 3.3.2) were compared to the finalized lists of both references and also to the Jagannathans and colleagues' variant list. As coordinates of 'de novo' variants were different between reference versions, the positions needed to be converted to each other for the comparisons. This conversion was performed with the Remap (see Appendix Table A.A.1 for the version) algorithm provided by NCBI.

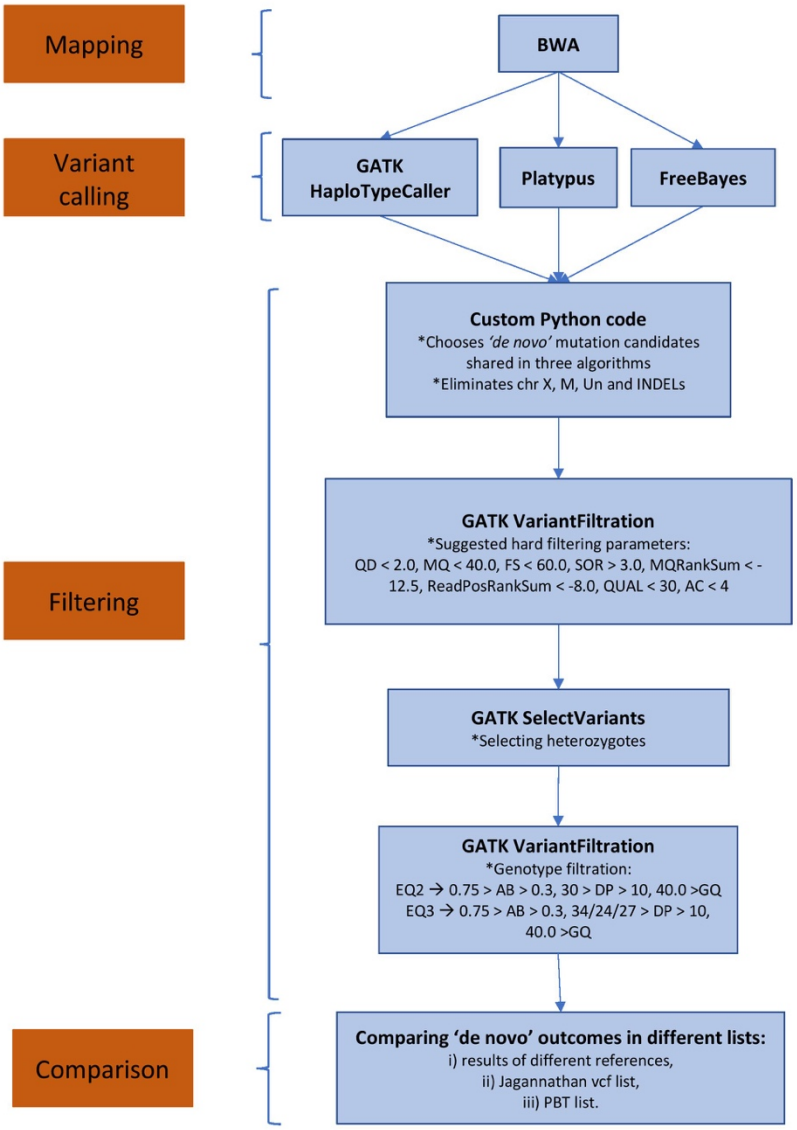


**Figure 3.1:** Workflow for detecting 'de novo' mutation candidates. There were four main parts which were mapping, variant calling, filtering and comparison with different variant lists. Main functions for filtering and their details are given in blue boxes.

### 3.4. Laboratory validation of EquCab2 candidates

All 'de novo' candidates predicted were checked in the IGV Genome Browser (see Table A.A.1). Based on this visual check, the most promising 50 'de novo' candidates were chosen in total for laboratory validation by Sanger resequencing. The panel consisted of 20 'de novo' variants in the Lipizzan trio (%39 of all Lipizzaner 'de novo' candidates), 14 in the Noriker trio (%24 of all Noriker 'de novo' candidates) and 16 in the Haflinger trio (%27 of aall Haflinger candidates). Validation candidates are given in Table A.C.2 in Appendix C.

For each 'de novo' variant, flanking primers were designed for Sanger resequencing. Primers flanking the selected variants were designed using the web tool Primer3 (see Table A.A.1). Primer information is given in Appendix C, Table A.C.2.

Laboratory validation were made by Doris Rigler and Eva Michaelis at Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna. Briefly, PCR amplification was performed using genomic DNA from each member of the respective trio (mother, father, son) as template. PCR reactions were carried out in a 20 µl volume containing 2µl genomic DNA (5 – 20 ng/µl), 0.5 µM of each primer, 1.5 mM MgCl and 1 x PCR buffer, 200 µM each dNTP and 0.1 U Taq DNA polymerase (Agrobiogen). The DNA was initially denatured at 95°C for 5 min, followed by 35 cycles of 30 s at 95°C, 30 s at annealing temperature (Supplementary Table S2) and 40 s at 72°C. After the 35 cycles, a final extension for 4 min at 72°C was performed. PCR products were visualized on a 2% agarose gel and purified using a QIAquick PCR purification kit. The concentration of the products was checked on a 2% agarose gel using the DNA ladder and then sent for Sanger sequencing to LGC genomics®. Results were visualized using the program Codon Code Aligner (see Table A.A.1). Laboratory validation were made by Doris Rigler and Eva Michaelis.

### 3.5. Mutation rate estimation

The number of truly validated 'de novo' mutations in EquCab2 mapped data were used to calculate an estimate of the mutation rate in one generation. To calculate the rate, the number of true 'de novo' mutations for each trio was divided by length of genome that was covered with the given thresholds in that trio (Data Analysis Part 3) as seen in the Equation 1. Overlapping genome sizes (bp) for each trio were calculated by multiplying overlapping genome percentages with two times of autosomal genome length (2,242,939,370 bp in EquCab2).

$$\text{Mutation rate in one generation} = \frac{\#\ of\ true\ de\ novo\ mutations}{\%\ overlapping\ genome\ size \times genome\ length \times 2} \quad (1)$$

### 3.6. Estimation of demographic history and runs of homozygosity

Pairwise Sequentially Markovian Coalescent (PSMC) analysis (Li & Durbin, 2011) was applied to estimate the demographic history of the breeds investigated. First, diploid consensus sequences were obtained from filtered (quality filtering and duplicate removal) versions of the BAM files (Data Analysis Part 2) mapped to the EquCab3 reference. SAMtools' 'mpileup' function with -C50 parameter and Bcftools (see Appendix A) were used to obtain these consensus sequences. While creating consensus sequences, a depth filter was applied to each trio again. The lower and upper thresholds were 10x to 2 times the mean trio coverage for each individual, respectively. In the next step, input files for the PSMC analysis were prepared with the 'q2psmcfa' function. The analysis was made with the suggested simulation parameters: "-N25 -t15 -r5 -p '4+25*2+4+6'", from the previous horse data analysis by Orlando and colleagues. Then, plots were drawn in R using the suggested generation time (8 years) and horse mutation rate (7.242e-09) from Orlando et al. (Orlando et al., 2013).

To estimate runs of homozygosity (ROH), variant files produced from EquCab3-mapped data by GATK HaplotypeCaller were used. Only the variants on the autosomal chromosomes were used in the analysis. ROHs were identified by PLINK (Purcell et al., 2007) (see Appendix A) with parameters: "--homozyg --homozyg-kb 500 --homozyg-snp 50 --homozyg-window-snp 50 --homozyg-window-het 3 --homozyg-window-missing 3 --homozyg-window-threshold 0.05 --homozyg-density 50".

# CHAPTER 4

# 4.RESULTS

## 4.1.Primary data analysis and quality-based trimming

The total number of paired-end reads of each individual's raw data is given in Table 4.1. Whereas Haflinger BW-356 had the highest number of total reads in the data, Haflinger BW-357 had the lowest. The quality of raw reads was checked with FastQC and visualized with MultiQC. In Figure 4.1, the number of reads in different lanes are shown. Most of the Noriker and Haflinger lanes had a high amount of duplicates.

**Table 4.1:** The total number of reads in the raw fastq files for each individual. The same number of reads exist for both pairs.

| Sample ID | Pedigree | Number of raw PE reads |
|---|---|---|
| **111** | Lip. Son | 317,112,113 |
| **113** | Lip. Dad | 325,516,763 |
| **166** | Lip. Mom | 298,406,535 |
| **BW-352** | Nor. Dad | 365,143,936 |
| **BW-353** | Nor. Mom | 298,670,837 |
| **BW-354** | Nor. Son | 313,454,280 |
| **BW-355** | Haf. Dad | 313,514,763 |
| **BW-356** | Haf. Mom | 451,365,261 |
| **BW-357** | Haf. Son | 296,405,435 |

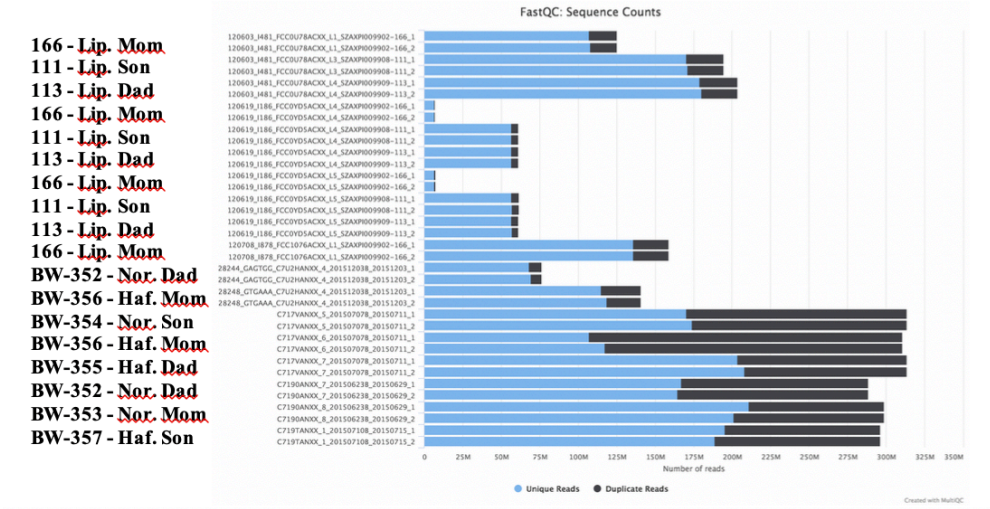


**Figure 4.1:** The number of paired-end reads in each lane. In the leftmost column, individual IDs and pedigree information for the lane are given. The proportion of duplicate (black) and unique (blue) reads for each lane are shown.

The number of reads and the mean Phred scores of different lanes are shown in Figure 4.2. While most of the reads increased for Noriker and Haflinger trios have very high mean Phred scores, in the Lipizzaner lanes the mean Phred score has a mode around 38 with few reads having qualities above 38.

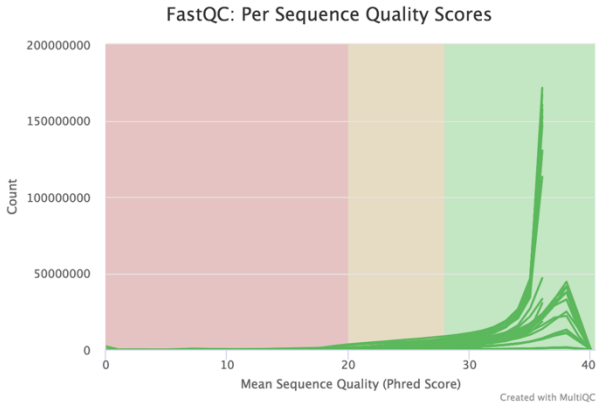


**Figure 4.2:** Mean Phred scores vs the number of reads of each lane. Most of the reads were above the mean Phred score of 30.

Figure 4.3 shows the distribution of each lane's mean Phred score at each nucleotide position. In the untrimmed data, the Lipizzaner parents' (113 and 166) reads had low Phred scores. Trimming parameters were thus chosen based on these low-quality lanes. After trimming, Phred scores were all over 30, throughout the read length. Length of the Lipizzaner reads were cut down to 81-86 base pairs (bp) from 90 bp, while Noriker and Haflinger reads were shortened to 99-113 bp from 120-125 bp.

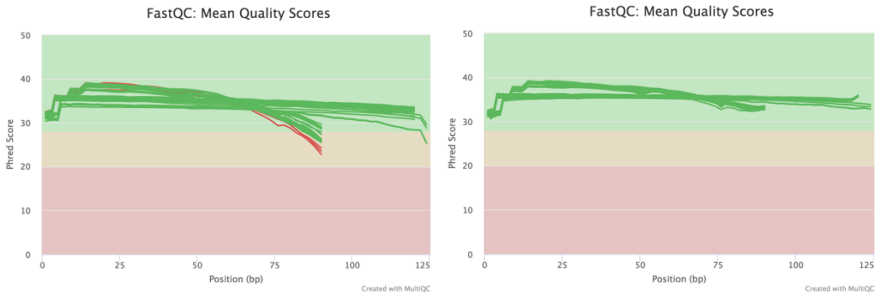


**Figure 4.3:** Mean Phred scores of each lane before (left) and after (right) trimming. After the trimming, reads with Phred score below 30 (red lines) for a position disappeared from data.

## 4.2.Mapping to the reference

The trimmed reads of each individual were mapped to the EquCab2 and EquCab3 reference genomes with BWA software. Different lanes that belonged to the same individual were merged into a single file. After that, the mapping files were filtered in order to remove unmapped reads and eliminate reads that were not properly paired and/or duplicated. In Table 4.2, the number of mapped reads in the raw data and in the filtered BAM files (i.e. properly paired and unique reads) are shown.

**Table 4.2:** The total number of reads for each individual, including both forward and reverse reads. BW-356 had the highest number of reads in the raw file, but after filtering this decreased dramatically, similar to the others. In the filtered data, the Lipizzan trio had the highest number of mapped reads.

| Sample ID | Pedigree info | Total # of reads in raw files | Total # of mapped reads in filtered files (EQ2/EQ3) |
|---|---|---|---|
| **111** | Lip. Son | 623366802 | 484898968/483215006 |
| **113** | Lip. Dad | 638708842 | 508612572/506977451 |
| **166** | Lip. Mom | 589312668 | 462960147/462505385 |
| **BW-352** | Nor. Dad | 699268926 | 303403168/303337894 |
| **BW-353** | Nor. Mom | 565506312 | 250130763/249834086 |
| **BW-354** | Nor. Son | 613144358 | 215904593/215968466 |
| **BW-355** | Haf. Dad | 613058170 | 291787760/291333599 |
| **BW-356** | Haf. Mom | 878454846 | 300332316/300064570 |
| **BW-357** | Haf. Son | 554469431 | 256835380/256756131 |

Mapping percentages of raw reads for different lanes are given in Table 4.3. Lipizzaners had the highest mapping rates (and properly paired read percentages) in both reference genomes while Norikers have the lowest.

Insert sizes of the final BAM files that were mapped to the EquCab3 reference were also calculated: Mean insert sizes were found as 475 bp in Lipizzaner, 308 bp in Noriker and 317 bp in Haflinger, and their medians were found as 476 bp, 306 bp and 316 bp respectively, in the finalized bam files.

### 4.3.Mapping coverage on autosomal chromosomes

Mean coverages on autosomal chromosomes of each individual are given in Table 4.4. All calculations were made for both reference genomes (EquCab 2 and 3). The Lipizzaner trio had the highest mean coverages, while Noriker individuals had the lowest.

Overlapping genomic regions which are in between threshold coverage values were calculated. The low and high thresholds were selected as 10 and 30, respectively, in EquCab2. In EquCab3, the low and high coverage thresholds were selected as 10 and twice the breed mean coverage, respectively. The lowest genomic size lying between these thresholds was calculated in the Noriker son (BW-354), whereas the highest was calculated in the Lipizzaner dad (113) for both references (see Table 4.4). Therefore, the Noriker trio had the lowest proportion of regions in the genome which were between the threshold values in all three trio-members and the Lipizzan trio had the highest proportion (Bold boxes in Table 4.4).

**Table 4.3:** Percentages of mapping and properly paired reads in each lane. Percentages were higher in the EquCab3 mapped data.

| Sample ID | Pedigree info | Lane | % mapping (EQ2, EQ3) | % prop. paired (EQ2, EQ3) |
|---|---|---|---|---|
| **111** | Lip. Son | 120619 I186 FCC0YD5ACXX L4 SZA XPI009908 | 97.63, 98.04 | 94.53, 95.10 |
| **111** | Lip. Son | 120619_I186_FCC0YD5ACXX_L5_SZA XPI009908 | 97.72, 98.13 | 94.69, 95.27 |
| **111** | Lip. Son | 120603_I481_FCC0U78ACXX_L3_SZAX PI009908 | 98.04, 98.45 | 95.23, 95.82 |
| **113** | Lip. Dad | 120619_I186_FCC0YD5ACXX_L4_SZA XPI009909 | 97.68, 98.11 | 94.46, 95.07 |
| **113** | Lip. Dad | 120619_I186_FCC0YD5ACXX_L5_SZA XPI009909 | 97.76, 98.20 | 94.61, 95.23 |
| **113** | Lip. Dad | 120603_I481_FCC0U78ACXX_L4_SZAX PI009909 | 97.66, 98.09 | 94.42, 95.03 |
| **166** | Lip. Mom | 120619_I186_FCC0YD5ACXX_L4_SZA XPI009902 | 97.66, 98.10 | 93.92, 94.45 |
| **166** | Lip. Mom | 120708_I878_FCC1076ACXX_L1_SZAX PI009902 | 98.43, 98.89 | 95.30, 95.87 |
| **166** | Lip. Mom | 120603 I481 FCC0U78ACXX L1 SZAX PI009902 | 98.12, 98.58 | 94.68, 95.25 |
| **166** | Lip. Mom | 120619_I186_FCC0YD5ACXX_L5_SZA XPI009902 | 97.76, 98.20 | 94.10, 94.62 |
| **BW-352** | Nor. Dad | 28244_GAGTGG_C7U2HANXX_4_2015 1203B_20151203 | 93.75, 94.15 | 87.51, 87.92 |
| **BW-352** | Nor. Dad | C7190ANXX_7_20150623B_20150629 | 94.00, 94.57 | 85.66, 86.21 |
| **BW-353** | Nor. Mom | C7190ANXX_8_20150623B_20150629 | 93.18, 93.70 | 85.57, 86.09 |
| **BW-354** | Nor. Son | C717VANXX_5_20150707B_20150711 | 95.12, 95.67 | 88.17, 88.72 |
| **BW-355** | Haf. Dad | C717VANXX_7_20150707B_20150711 | 94.96, 95.47 | 89.10, 89.62 |
| **BW-356** | Haf. Mom | C717VANXX_6_20150707B_20150711 | 94.88, 95.44 | 89.24, 89.87 |
| **BW-356** | Haf. Mom | 28248_GTGAAA_C7U2HANXX_4_2015 1203B_20151203 | 94.20, 94.64 | 88.50, 88.99 |
| **BW-357** | Haf. Son | C719TANXX_1_20150710B_20150715 | 95.06, 95.58 | 88.23, 88.76 |

**Table 4.4:** Mean coverages and the percentages among genome in between depth thresholds. While 10-30 depth thresholds were used in the EquCab2 data, 10 to 2 times breeds mean coverages of each trio were used in the EquCab3 mapped data.

| Sample ID | Pedigree info | Mean Genome Coverage (EquCab2) | % of genome between 10-30 DP (EquCab2) | Overlapping genome size in 10-30 DP (EquCab2) | Mean Genome Coverage (EquCab3) | % of genome between 10-34/24/27 DP (EquCab3) | Overlapping genome size in 10-34/24/27 DP (EquCab3) |
|---|---|---|---|---|---|---|---|
| **111** | Lip. Son | 17.32 | 88.17 | | 17.16 | 88.11 | |
| **113** | Lip. Dad | 18.05 | 88.25 | **71.6** | 17.89 | 88.86 | **73.96** |
| **166** | Lip. Mom | 15.96 | 81.68 | | 15.87 | 81.86 | |
| **BW-352** | Nor. Dad | 14.23 | 83.62 | | 14.16 | 80.60 | |
| **BW-353** | Nor. Mom | 11.36 | 66.01 | **37.43** | 11.29 | 64.89 | **35.53** |
| **BW-354** | Nor. Son | 10.21 | 56.02 | | 10.17 | 55.30 | |
| **BW-355** | Haf. Dad | 13.73 | 81.34 | | 13.65 | 80.02 | |
| **BW-356** | Haf. Mom | 13.92 | 81.99 | **56.46** | 13.83 | 80.54 | **54.96** |
| **BW-357** | Haf. Son | 12.16 | 71.949 | | 12.11 | 71.02 | |

## 4.4. 'de novo' mutation candidates

For detecting 'de novo' mutation candidates, the three main steps are mapping, variant calling, and filtering. These are followed by a comparison of the resulting variation lists ('de novo' candidate lists based on EquCab2 and EquCab3, Jagannathan's variation panel, PBT 'de novo' candidates). Figure 4.4 shows the whole 'de novo' candidate detection workflow and the number of variants after each step. The same workflow was applied to both EquCab2 and EquCab3 mapped data, and both results were given in the figure respectively in each box.

### *4.4.1. Variant calls on EquCab2 and EquCab3*

Three different algorithms were used to call variants in trios as shown in the variant calling part of Figure 4.4. GATK found the highest number of variants in all trios for both reference genomes, but the number of variants found in the updated horse genome release

EquCab3 was lower. The initial numbers given at the output of the variant calling step in Figure 4.4 include both SNVs and indels.

### *4.4.2. Filtering variants to detect 'de novo' mutation candidates*

The filtering of the variants was started by finding 'de novo' candidates with a Python script (see Materials and Methods part 3.3.1) and eliminating variant calls on chromosomes X, M and Un, and also insertions and deletions (INDELs). At this point, the number of 'de novo' mutation candidates were highest in the results of the Freebayes algorithm, among the three algorithms, for each trio and for both reference genomes (see Figure A.B.1 and Figure A.B.2 in Appendix B). In the next step, shared candidates were extracted between three algorithms and the highest number of common candidates were found between the GATK HaploTypeCaller and FreeBayes as seen in Appendix B, again for both reference genomes. On the other hand, the intersection of candidates between the three algorithms was highest in the Noriker trio and lowest in the Lipizzan trio.

Subsequently, variants were filtered based on different properties (heterozygosity, quality, etc.) as can be seen in the filtering part of Figure 4.4. At the end of the filtering, 51, 58 and 59 'de novo' candidates in EquCab2 mapped data, and 69, 45 and 55 candidates in EquCab3 mapped data were predicted in the Lippizaner, Noriker and Haflinger trios, respectively. Because different depth thresholds were applied to the variants at the last step of filtering, the minimum and the maximum number of resulting 'de novo' candidates were different in over trios and reference genomes. Detailed information of all 'de novo' candidates from the two reference genomes for the three trios is given in Table A.C.1 in Appendix C.

### *4.4.3. Comparing finalized candidates from different lists*

In the last step of the workflow (Figure 4.4), finalized candidate lists from EquCab2 and EquCab3 and the common variant list from Jagannathan and colleagues (Jagannathan et al., 2018) which was created by 88 horses from EquCab3 reference genome were compared. Figure 4.5 shows the results of the comparison. Detailed information on 'de novo' SNPs is provided in the Table A.C.1. Candidates were also compared with the results from 'de novo' mutation detection algorithm PhaseByTransmission in GATK. Finalized 'de novo' candidates were examined in the IGV Genome Browser and chosen ones validated in the laboratory (see the subsection 4.5 titled *Laboratory Validation of 'de novo' Candidates* below).

#### *a) Comparing to Jagannathan's common variant panel*

38, 27, and 30 shared 'de novo' mutations were found between EquCab2 and EquCab3, respectively in Lipizzan, Noriker and Haflinger trios. Most of these 'de novo' mutation candidates were also found in the Jagannathan variation panel as seen in Figure 4.5. In particular, only 14 Lipizzaner, 5 Noriker, and 12 Haflinger 'de novo' variants did not appear in Jagannathan's horse population variation list. Positions and detailed comparison results of all these mutations were given in Table A.C.1.

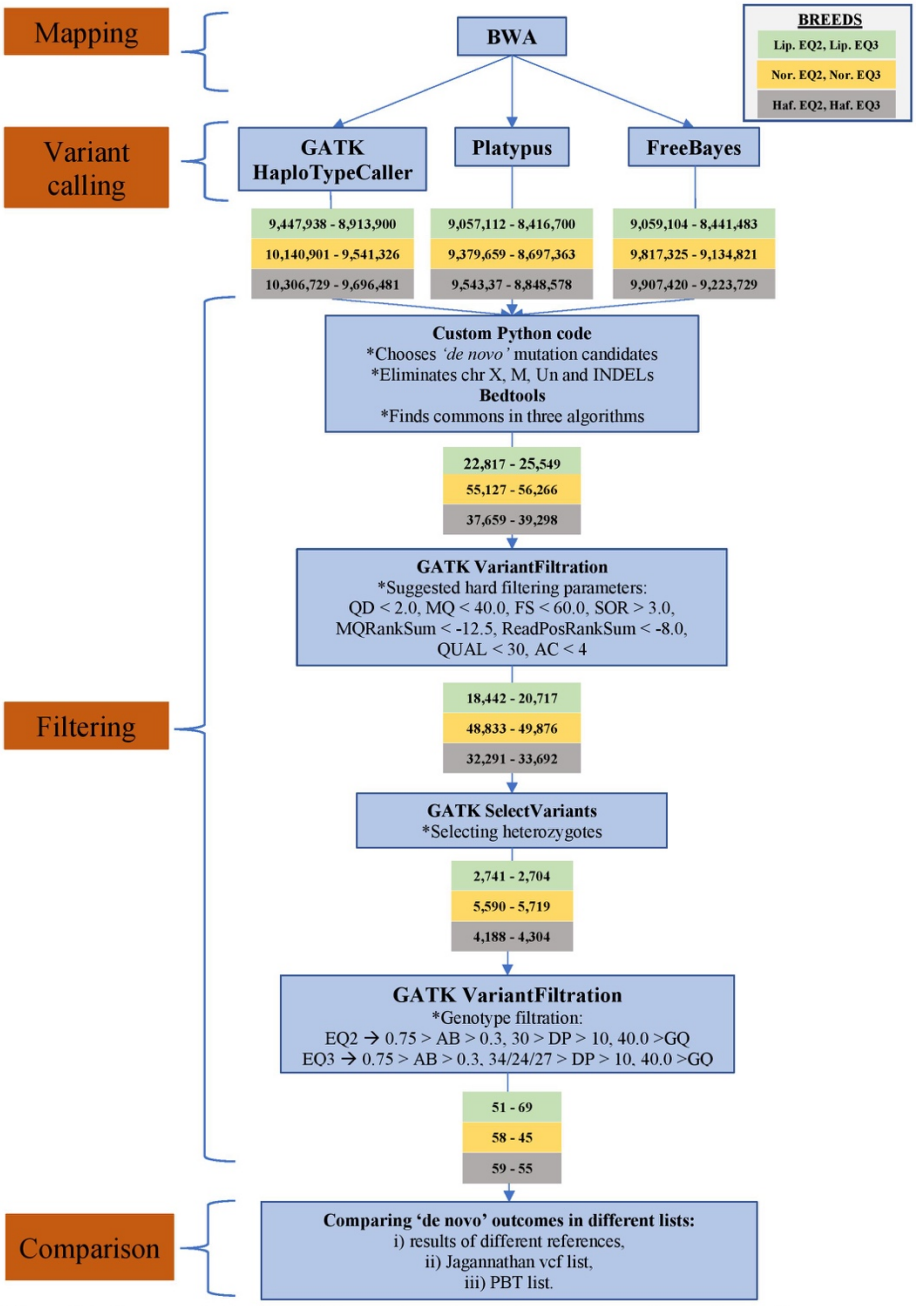


**Figure 4.4:** Shows analysis steps and the number of remaining mutations after each step. Blue boxes show the steps of the analysis, whereas green, yellow and grey boxes show the number of mutations in different breeds and different references after each step.

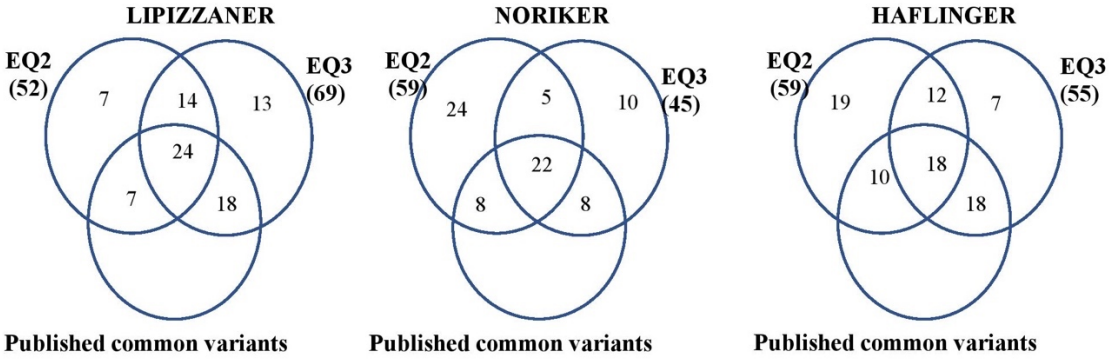


**Figure 4.5:** Comparison of finalized lists from the two reference genomes and published common variants. The highest number of shared mutations between the EquCab2 and EquCab3 genomes were found in the Lipizzan trio. Total number of mutations in a set are given below the name of the set.

*b) Comparing to the PhaseByTransmission result*

'De novo' candidates in the Lippizaner trio were also predicted using the GATK PhaseByTransmission algorithm. Only 5 different candidates were detected at the intersection of the three different variant calling algorithms (GATK HaplotypeCaller, Freebayes, Platypus). These five variants were heterozygous in the offspring and homozygous for the identical allele in both parents. Positions and coverages of these five 'de novo' candidates are shown in Appendix C, Table A.C.2. Neither of these candidates were found in the EquCab2 or EquCab3 'de novo' candidates generated in our custom workflow above, but three of them were found among Jagannathan's common variants.

## 4.5. Laboratory validation of 'de novo' candidates

Candidate 'de novo' mutations revealed after the filtering steps, as explained above, from EquCab2 mapped data (51 Lipizzan-, 58 Noriker- and 59 Haflinger-variants) were visually checked in the IGV Genome Browser. Candidates that were located near other variants were eliminated and 50 candidate variants (20 in Lipizzaner, 14 in Noriker, 16 in Haflinger) were selected for independent lab validation. Detailed information of the validated variants is given in Table A.C.3, in Appendix C. For the chosen variants, locus-specific flanking primers were designed and a PCR product was amplified in mother (M), father (F) and son (S) and these were sequenced with Sanger technology.

The results of validation are shown in Table 4.5. 13 out of the 20 the Lipizzaner (around %65 of all tested candidates), 3 out of 14 the Noriker (around %21 of all tested candidates) and 5 out of 16 the Haflinger (around %31 of all tested candidates) candidates were successfully validated. In 5 Lipizzaner, 4 Noriker and 4 Haflinger candidates the putative 'de novo' allele was also detected in one of the parents. Some of the candidates could not be validated at all because Sanger sequencing did not produce good results in those positions (given in the *Missing* row in Table 4.5). Finally, some of the positions did not show the 'de novo' allele as observed in the NGS data. These are called 'No SNV' in

Table 4.5. IGV screenshots and Sanger validation results of a true ‘de novo’ mutation, a position that was not found as a variation in Sanger sequencing (i.e. ‘no SNV’) and a ‘de novo’ candidate that was also detected in the parents are, both shown in Appendix D.

**Table 4.5:** Results of laboratory validation. While some of the SNPs could not be detected with validation, some of them were also found in the parents. The highest proportion of true ‘de novo’ mutations were detected in the Lipizzan trio.

| | **Lippizaner** | **Noriker** | **Haflinger** |
|---|---|---|---|
| **Total number of validated SNVs** | 20 | 14 | 16 |
| **Missing (bad Sanger sequence)** | 1 | 1 | 2 |
| **SNV found also in parents** | 5 | 4 | 4 |
| **No SNV** | 1 | 6 | 5 |
| **Validated ‘*de novo’* variant** | 13 | 3 | 5 |
| **% of truly validated candidates** | ~%65 | ~%21 | ~%31 |

Figure 4.6 shows the distribution of validation results and the comparison of three variant lists in addition to Figure 4.5. The 5 Lipizzaner, 2 Noriker and 3 Haflinger ‘de novo’ SNPs that were found also in the parents in the validation result were also found in the Jagannathan's variants. Besides, except one in Lipizzaner, all of the true positive ‘de novo’ mutations were detected in both EquCab2 and EquCab3 mapped data (see also Figure 4.6).

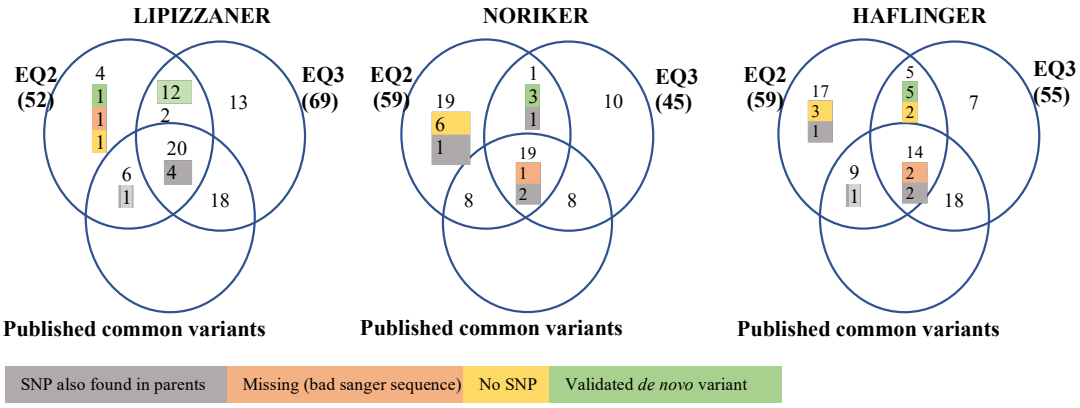


**Figure 4.6:** Comparison of three variant lists including validation results of EquCab2 mapped data. Grey label refers to SNPs which also determined in one of the parents in validation, orange label refers to bad Sanger sequenced positions, yellow label refers to variants not validated as SNP in Sanger results and green labels refers to validated 'de novo' variants. Except one, all true 'de novo' mutations are in the intersection of EquCab2 and EquCab3 mapped data and they are not seen in the Jagannathan's variants.

## 4.6. 'de novo' mutation rate

The number of true 'de novo' mutations detected on the EquCab2 mapped data was used to calculate mutation rate in one generation (see Materials and Methods, part 6). To calculate mutation rates overlapping genome percentages (71.6 in Lipizzaner, 37.43 in Noriker and 56.46 in Haflinger) in the 10 to 30 depth threshold values were used as genome sizes and these revealed rates of ~ 4e-09 bp/gen, ~1.7e-09 bp/gen and ~1.9e-09 bp/gen bp/gen in Lippizaner, Noriker and Haflinger trios, respectively. Detailed calculation steps are given in the Materials and Methods.

## 4.7. Estimation of demographic history and runs of homozygosity

To estimate the demographic history of the three populations, a Pairwise Sequentially Markovian Coalescent (PSMC) analysis was performed. The resulting PSMC graph is shown in Figure 4.7. All nine horses from three different breeds showed a similar pattern. There was a dramatic decrease in the effective population size around 500,000 years before present and the effective population size reached its maximum level around 120,000 years ago. The logarithmic version of the PSMC graph is given in Appendix E, Figure A.E.1.

The histograms showing the distribution of homozygous runs over lengths of runs are given in Figure 4.8. As evident from the figure, and as is expected, the number of runs decreased when the length of the run increased. The three Lippizan horses had the highest mean number of homozygous runs for each histogram bin, whereas the Haflingers had the

lowest. In general, the distribution of the Haflinger and Noriker trios' runs of homozygosity were closer to each other, when compared with the Lippizaner trio.

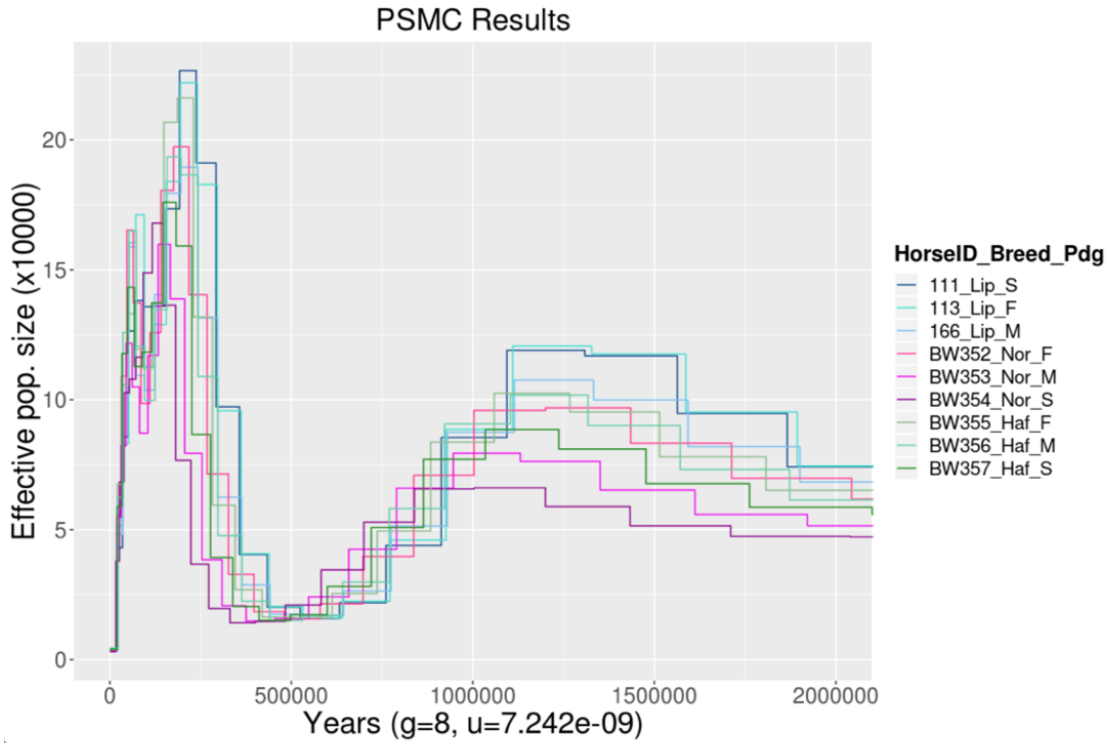


**Figure 4.7:** PSMC results for nine horses from three breeds as a function of time before present and effective population size. E.g., 2e+06 in the x-axis is equal to 2 million years before present.

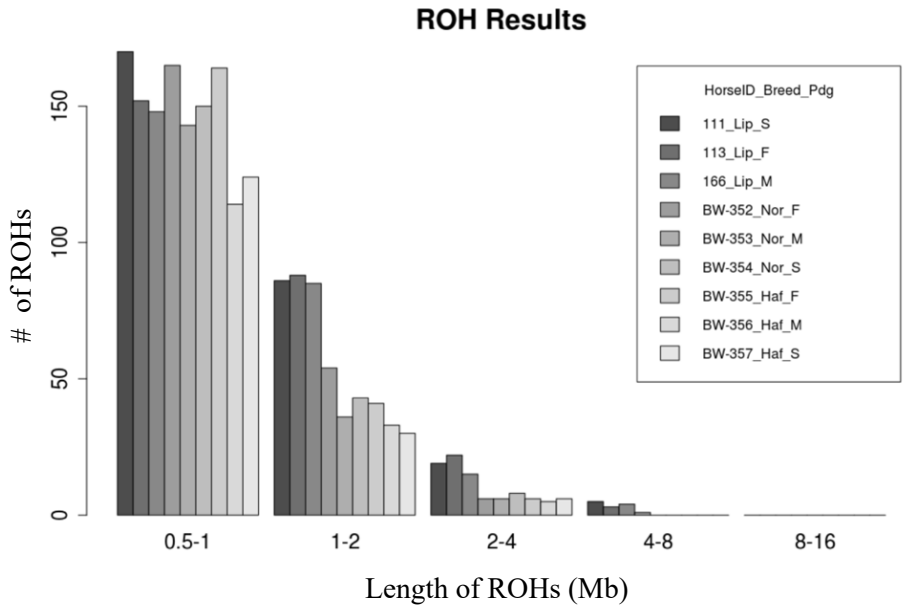


**Figure 4.8:** Distribution of homozygous runs for different length intervals. The x-axis refers to length of homozygous runs in megabases, the y-axis shows number of homozygous runs.

# CHAPTER 5

# 5. DISCUSSION

In this thesis, we analyzed whole genome sequencing data of three horse trios. We designed a workflow to determine 'de novo' mutations accurately. We then calculated the mutation rate per generation with these 'de novo' mutations and compared this to the mutation rates from previous studies. Additionally, we made estimations regarding the demographic history of the three horse breeds by using PSMC and ROH analysis and compared results with previous studies.

## 5.1. Data, mapping, coverages and variant calling in different reference genomes

Some of the NGS sequencing lanes in the Noriker and Haflinger samples, produced by Vienna BioCenter (CSF) in 2015, had high numbers of duplicate reads. There are four main reasons for a duplicated read i) biological duplication (repeat regions), ii) PCR duplications, iii) optical duplicates which are the duplicated reads occur by accepting large clusters as two different clusters by mistake or by a local re-clustering of the original library polymer during bridge amplification, iv) ExAmp duplicates (Hadfield, 2016). Any of these can cause duplicate reads in Illumina sequencing. In our case, the main reason for having a high number of duplicates in some lanes in the Noriker and Haflinger data is assumed to be optical duplication, due to the sequencer used; HiSeq 2500 run mode HiSeq SBS V4. It has been suggested that it is more likely to have optical duplicates with this particular device and run mode (Hadfield, 2016).

When we compared mapping results between the EquCab2 and EquCab3 references, while the number of mapped reads after filtering were lower in EquCab3 mappings (Table 3-2), mapping percentages were increased in EquCab3 compared to EquCab2 mappings (Table 3-3). The increase in mapping ratio for the updated reference genome was also shown by Kalbfleisch and colleagues (Kalbfleisch et al., 2018). Besides mapping percentages, the number of properly paired reads also increased in the EquCab3 mappings. When we called variants with three different algorithms, we observed a decrease in the number of variant calls in the EquCab3 mapped data. This decrease in the number of

variants implies that false positive variants called in the earlier genome version (EquCab2) are now eliminated, as suggested in Guo et al., 2017.

The mean mapping coverage of each individual and overlapping genomic regions (i.e. regions where all members of a trio have coverage in between given thresholds) in each trio also differed between reference genomes (Table 4.4). The difference in the percentage of the overlapping genomic regions was mainly due to different covering thresholds being applied for the two different reference genomes (see 3.2.3). These changes in the threshold levels resulted in a wider depth range in the Lipizzaner trio in EquCab3 (threshold is 10 to 34 instead of 10 to 30 in EquCab2) and accordingly a larger percentage of the genome falling into this category in EquCab3. In contrast, a narrower range in EquCab3 in the Noriker (threshold is 10 to 24 instead of 10 to 30) and Haflinger (threshold is 10 to 27 instead of 10 to 30) lead to a decrease in percentages.

## 5.2.Variant calling with different algorithms and filtering

We performed variant calling with three different algorithms (GATK HC, Platypus, Freebayes). GATK HC revealed the highest number of variants among them as seen in Figure 4.4. In the first step of filtering, we chose ‘de novo’ mutation candidates which show patterns as given in Table 2.2. Most variants were eliminated in this step as shown in Figure A.B.1 and A.B.2. This is expected because most variants are not expected to be ‘de novo’ candidates. In the next step, by eliminating the X, and Un chromosomes and the indels we lost almost half of the ‘de novo’ candidates. After this elimination, remaining candidates from three different algorithms were compared with each other to find shared ones. We observed that FreeBayes found the highest number of unique candidates which were not detected by any other algorithms, in the Noriker (83,593 in EQ2, 83,857 in EQ3) and Haflinger (56,279 in EQ2, 57,822 in EQ3) breeds (see Figures A.B.1 and A.B.2). GATK HC found the second-highest number of unique candidates in these trios. One reason for having the higher number of unique mutations called with different algorithms in Noriker and Haflinger could be the low sequencing coverage of some individuals in these trios. Tatsumoto et al., 2017 also suggests that false-positive variant prediction is more likely to occur in low coverage data. We also compared the intersection of the number of candidates detected by different algorithms pairwise and found the highest number of intersecting ‘de novo’ candidates among Freebayes and GATK HC, for both reference genomes and for each trio. Because this intersection did not include the candidates shared by all three of the algorithms, we speculate that most of these were false-positives, given that Platypus has a lower false-positive rate. We also found that the Lipizzan trio had the lowest number of mutations in the pairwise intersections of different algorithms as well as at the intersection of all three algorithms. This is consistent with the expectation of fewer false positives with the higher coverage of the Lipizzan data; fewer candidates that are more accurate are observed.

GATK Hard Filtering parameters were suggested by Summa et al. (De Summa et al., 2017) to detect the qualified variations among all variations. Around 15 % of the

candidates were filtered out in each trio and each reference with hard filtering. Selecting heterozygote 'de novo' candidates among all patterns (Table 2.2) also caused a dramatic decrease in the number of candidate 'de novo' mutations. To see a 'de novo' mutation having a homozygous allele pair in the offspring (see first four rows in Table 2.2), one of the parents should carry the same homozygous allele with the offspring and the other parent should be heterozygous at that position. This is a vanishingly low probability because it requires independent mutations in the offspring and one of the parents, or a back mutation (a mutation at the same position as a previous mutation that reverts the change of the former back to the reference allele), at that same position. Likewise, when the offspring is heterozygous (0/1) at a position, having homozygous alternative alleles (1/1) in the parents also means that a back mutation occurred in the position (see the last row of Table 2-2). Due to the extremely low probabilities of the aforementioned events, we speculate that only candidate 'de novo' patterns where the offspring is heterozygous (0/1), while the parents carry the homozygous reference allele (0/0) (see the fifth row of Table 2-2), have a chance to be a real 'de novo' mutation. We therefore focused on this pattern, applied filtering to choose this pattern among all 'de novo' patterns, and saw a dramatic decrease in the number of candidates. By eliminating the other patterns we speculate that we must have eliminated of most of the false positive candidates (e.g. due to sequencing errors) in our data.

We finally filtered candidates for allele balance and depth, as suggested in several studies (Jónsson et al., 2017; Tatsumoto et al., 2017). In EquCab2 data, we filtered for 'de novo' candidates with 10 to 30 read depth values and we got the lowest number of candidates in Lipizzaner trio (51), whereas higher numbers in Noriker (58) and Haflinger (59) trios. Because the lowest overlapping regions between 10 to 30 depth values were found in the Noriker, followed by the Haflinger trios, we speculate that a higher number of false positives in the 'de novo' candidate list of Noriker and Haflinger trios caused this observation. We therefore changed the depth thresholds from 10 to two times mean coverages of the respective trios when analyzing the EquCab3 mapped data. Because mean coverage is lower in Noriker and Haflinger trios, their upper thresholds were thus also lower (24 for Noriker, 27 for Haflinger), while Lippizzaner had a higher upper threshold (34). These changes in the threshold caused an increase in the number of 'de novo' candidates in the Lipizzaner (69) list and decrease in Noriker (45) and Haflinger (55) trios. This shows that setting the upper limit as a function of the mean lowers the probability of false positives (see Figure 4.5) especially in low covered trios (Noriker and Haflinger). Keeping depth filtering as a function of mean was also suggested by Li and Wren (Li & Wren, 2014) When we compared the number of 'de novo' candidates with the numbers in other mammalian studies (Conrad et al., 2011; Jónsson et al., 2017; Tatsumoto et al., 2017), we saw that they were comparable.

### 5.3.Interpretation of 'de novo' candidates and Sanger validation results

We compared the candidate lists from EquCab2 and EquCab3 mapped data and found the highest number of shared mutations in Lipizzaner (38) and the lowest in Noriker (27)

(Figure 4.5). We expected that intersecting candidates on two independent callings using different reference genomes were the best candidates for true 'de novo' mutations. Besides, to get a feeling how many of false positive candidates are still in our list, we compared the candidates to a recently published horse variation dataset given in Jagannathan et al. (Figure 4.5) (Jagannathan et al., 2018). For our purposes, the Jagannathan panel serves as a database for standing variation in the Central European horse population. We found a remarkable overlap especially in the Noriker and Haflinger 'de novos' as shown in the Figure 4.5. Possible explanations for this observation are sequencing errors or low coverage sequencing of parents' data. Because heterozygosity in one parent was not detected with Illumina sequencing, some variants falsely give a 'de novo' mutation pattern after filtering. We saw this ratio was higher in the Noriker (22 out of the 27 'denovo' variants also in Jagannathan dataset) data than in the Lipizzaner (24 out of 38) and Haflinger (18 out of 30) data. These results support our hypothesis of a higher amount of false positives in the lower coverage Noriker trio.

We performed Sanger resequencing for a selected set of 'de novo' mutations after checking all candidates in the IGV Genome Browser for both parents and offspring. Results of laboratory validation allowed us, on the one hand, to infer the validity and true number of our 'de novo' predictions, and on the other hand, they illuminate possible scenarios leading to false positive candidates. For example, i) Some could be sequencing errors in the offspring and we called them 'no SNP' in the validation results. ii) The putative 'de novo' allele was also detected in one of the parents with Sanger sequencing. These positions were most probably not called with Illumina sequencing because of the low coverage. We called these 'SNP also in parent'. iii) 'de novo' candidates that could not be confirmed with Sanger sequences, due to inconclusive Sanger results, were classified as 'bad Sanger sequence'. Such errors occur when the position is in for example in a duplicated regions in the genome, as suggested in the Tatsumoto et al., or the PCR amplicon for Sanger sequencing was not locus specific; iv) Finally, true 'de novo' mutations that were validated successfully and the 'de novo' allele was detected only in the offspring by both NGS and Sanger sequencing.

Validation results (Table 4.5) were consistent with the data quality. Ratio of truly validated mutations was 13/20, 3/14 and 5/16 in Lipizzaner, Noriker and Haflinger trios respectively. The results clearly showed that, sufficient read coverage is essential to detect true positive mutations. We expected this result because we had the of the highest mean coverage in the Lipizzaner (~17x) and the lowest in the Noriker (~12x) data. Tatsumoto et al. suggested that even 30x and 60x coverages are not enough to detect all 'de novo' mutations in a trio data (Tatsumoto et al., 2017). Although our data was not nearly that deeply covered, with our custom filters we have reduced the number of false positives and detected a healthy number of 'de novo' mutations which are true positive.

We make two inferences from our results at this point: First, Lipizzaner data coverage is at the edge of the coverage necessary to reliably detect a sufficient number of true 'de novo' mutations. Second, our workflow still gives partially accurate results, in spite of the

low coverage. To sum up, having genomic data with at least 17x mean coverage and applying well-adjusted depth thresholds could be the most important steps in finding true 'de novo' mutations.

Number of candidates and true positives also correlate with the size of the genomic regions that were considered for analysis. The Lipizzaner trio had the largest genomic area (71,60% in EquCab2) within the thresholds, whereas Noriker had the lowest (37.43% in EquCab2). Based on this, we expect to find more candidates and true 'de novos' in Lipizzaner, and less in Noriker. Figure 5.1 shows the relation between scanned regions (1.605 x 109 Mb in Lipizzaner, 0.839 x 109Mb in Noriker and 1.266 x 109 Mb in Haflinger) and true positive mutations (13 in Lipizzaner, 3 in Noriker and 5 in Haflinger).

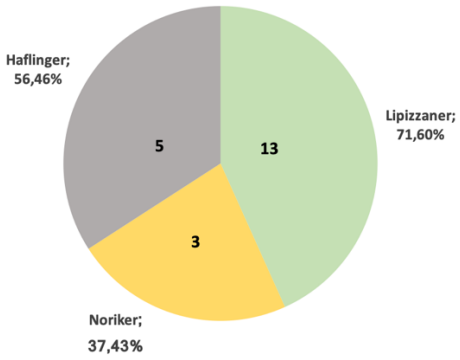


**Figure 5.1:** Percentages of overlapping genomic regions that were covered between 10-30 in EquCab2 in the respective trio and number of true 'de novo' mutations.

We also checked the distribution of the validation results among EquCab2, EquCab3 and the variant dataset from Jagannathan et. al. We found that all of true 'de novo' mutations, except one, in Lipizzaner were detected in both EquCab2 and EquCab3 mapped data but not in the Jagannathan variation panel, as expected (Figure 4.6). We found that many of the candidates which were classified as 'SNP also found in parents' were also in the Jagannathan data. Seeing these variants already in the Jagannathan data is further proof that Illumina sequencing missed heterozygous alleles in the parents due to very low sequencing coverage. None of the 'no SNP' positions were found in the Jagannathan's variation list. This was expected and is another confirmation that these positions are sequencing errors in the offspring.

## 5.4.Mutation rate differences

We calculated a 'de novo' mutation rate to be 4.05 per billion sites per generation for the Lippizaner, 1.79 for the Noriker, and 1.97 for the Haflinger trio, by dividing number of true positive 'de novo' mutations detected in each trio with 2 times overlapping genomic area of the respective trio in 10-30 coverage thresholds. We compared these ratios with human, chimpanzee and mouse rates which were calculated with the same equations.

When we compare these rates with the germ line ‘de novo’ mutation rate of human which is 1.29e-08 (Jónsson et al., 2017) our rates are almost 3 to 6 times smaller than the human rate. In the chimpanzee case the mutation rate was calculated as 1.48e-08 (Tatsumoto et al., 2017) and our Lippizzan rate was almost 4, Noriker rate was almost 9, and Haflinger rate was almost 8 times smaller than chimpanzee rate. When we compared the estimated horse rates with the mouse mutation rate estimates, ranging from 3.9e-09 (Lindsay et al., 2016) to 5.4e-09 bp/gen (Uchimura et al., 2015), our rates appeared comparable with them. Besides, when we compared the estimated rates with the mutation rate calculated for horses by phylogenetic distance method which is 7.2e-09 bp/gen (Orlando et al., 2013), we could say that we estimate the rate at least in the same decimals. One possible reason of the difference between phylogenetic and 'de novo' rates could be the false negative ‘de novo’ mutations that we missed during the analysis because of our data quality. However, other explanations are also possible (see Introduction).

When we compared different mutation rates between our three horse trios, we found the highest rates in the Lipizzaner whereas Noriker and Haflinger rates were lower and closer to each other. We speculate that the main reason for this difference is again the quality of the sequencing data. We had the highest coverage in the Lipizzaner trio, and thus the most accurate ‘de novo’ mutation rate is the one calculated from Lipizzaner data. Other possibilities of getting different rates could be fathers’ age (Francioli et al., 2014; Kong et al., 2012) and background effects differing between families (Conrad et al., 2011).

## 5.5.Estimations on population history

### *5.5.1. PSMC*

To estimate the demographic history of *Equus caballus*, we performed PSMC analysis. Results of the analysis revealed patterns similar to the ones presented by Orlando et al., 2013 and Schubert et al., 2014. The horse effective population size reached a minimum value between 100 to 130 thousand years (kyr) before present in our data. Orlando et al. found the same decrease 125 kyr before present, and they suggested that the Last Interglacial Period was the reason for this decrease. This decrease is followed by a peak in the effective population size in our findings, as well as in Orlando et al., 2013. This peak is thought to have developed due to the environmental changes after the Last Glacial Maximum (Lorenzen et al., 2011). However, although this timeline is consistent with our findings, there are still some differences in the timeline vs. effective population size patterns of our nine horses. One reason for this could be, again, the low mean coverage of our data. Nadachowska-Brzyska and colleagues have suggested that the PSMC method is not able to produce reliable models under 18x mean coverage, and our mean coverages in EquCab3 mapped data ranged between 10.17 and 17.89 (Nadachowska-Brzyska et al., 2016). The difference between mean coverages could also explain the differences in the models of different individuals. Despite these differences, we suggested that taking the regions in between 10 to 2 times mean coverage threshold (Li & Wren, 2014) provides

consistent pattern among individuals and with the other studies (Orlando et al., 2013; Schubert et al., 2014).

### *5.5.2. Runs of homozygosity analysis (ROH)*

We also performed a 'runs of homozygosity' analysis on our data to estimate population histories of the three different Austrian breeds in our study. Ceballos and colleagues suggested that inbred populations have a higher number of longer ROHs than outbred populations. They also suggested that if the population has a high number of shorter runs and along with some longer ROHs, it is possible that the population is both inbred and bottlenecked (Ceballos et al., 2018). When we interpreted our results from this point of view, we saw that the Lipizzaner trio is more inbred than the Noriker and Haflinger trios. Previous studies have found higher inbreeding in Austrian Lipizzaners with respect to other Lipizzaners (Grilz-Seger et al., 2019), and more inbreeding in Lipizzaners as compared to Norikers (Druml et al., 2009). Additionally, no appreciable difference between Lipizzaner and Haflinger populations in inbreeding was reported (Grilz-Seger et al., 2019).

Our ROH results which showed a low number of long ROHs (>4Mb) and high number of short ROHs (<2Mb) in Noriker and Haflinger trios, suggest less inbreeding in recent generations when compared with Lipizzaners (Ceballos et al., 2018). The conclusion of more inbreeding in Lipizzaner compared to Noriker is also supported by Druml et al., 2009. On the other hand, more inbreeding in Lipizzaner than Haflinger was not reported previously. This could be an effect of coverage differences in our data or might reflect variation within these breeds.

## 5.6.Future work and conclusion

We only validated candidate mutations from EquCab2 mapped data, but in addition we defined new possible candidate 'de novo' mutations called on EquCab3 only, which were not found in Jagannathan's data set (green labeled candidates in Appendix Table A.C.1). These 'de novo' candidates should be validated as well. While validating them in the future, besides taking the blood sample used for Illumina sequencing as a source for Sanger validation, another somatic tissue also could be used to distinguish true germ line 'de novo' mutations from somatic 'de novo' mutations. Only if the same 'de novo' mutations are validated in the other somatic cells, we can be definitely sure that they are germ line and not somatic cell mutations (Tatsumoto et al., 2017).

The statistical power of 'de novo' mutation detection workflows should also be measured in terms of false negative rate. To calculate false negative rate, we would need to estimate 'de novo' mutation candidates with another data set which have also validated true 'de novo' mutations. By trying our workflow on such a more deeply sequenced data we could have an estimate on false negative rate, and thus could obtain a more accurate mutation rate per generation for horse. For this aim we can analyze publicly available genomic

datasets, which include truly validated ‘de novo’ mutations (Francioli et al., 2014), in order to test its reliability. Another option could be analyzing the 'Genome in a Bottle' data set which includes deep coverage (~300x) human trio data (Zook et al., 2016), and we could compare verified ‘de novo’ mutations from this data set with ‘de novo’ mutations estimated from our workflow (Zook et al., 2018) and using the same coverages as in the horse data.

The demographic analysis could also be enriched, e.g. by calculating the inbreeding coefficient ($F_{ROH}$) from ROH data, by dividing average length of ROHs to autosomal genome size (McQuillan et al., 2008), and comparing these estimates with other studies’ outcomes (Druml et al., 2009; Grilz-Seger et al., 2019).

In conclusion, our data is at the very low-end coverage compared to many other studies for 'de novo' mutation estimation. Nevertheless, we were able to analyze this data to get as much information out of it as possible. We were able to detect a significant number of the true ‘de novo’ mutations from trio data with our workflow, and this was the achievement of our workflow for ‘de novo’ mutation detection. On the other hand, PSMC and ROH results were consistent with previous studies in most cases. Our study thus provided insight into the minimum data quality needed for such analyses.

## REFERENCES


Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, Vol. 17, p. 241. https://doi.org/10.1186/s13059-016-1110-1

Andrews, S. (2015). FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Institute*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/

Anthony, D. W. (2016). The Horse in Human History . Pita Kelekna . *Journal of Anthropological Research*, *66*(3), 401–403. https://doi.org/10.1086/jar.66.3.20798830

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Broman, K. W., & Weber, J. L. (1999). Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *American Journal of Human Genetics*, *65*(6), 1493–1500. https://doi.org/10.1086/302661

Campbell, C. D., & Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends in Genetics*, *29*(10), 575–584. https://doi.org/10.1016/j.tig.2013.04.005

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, *19*(4), 220–234. https://doi.org/10.1038/nrg.2017.109

Chen, H. (2015). Population genetic studies in the genomic sequencing era. *Zoological Research*, *36*(4), 223–232. https://doi.org/10.13918/j.issn.2095-8137.2015.4.223

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., … Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–714. https://doi.org/10.1038/ng.862

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., & Tommasi, S. (2017).

GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics*, *18*(Suppl 5). https://doi.org/10.1186/s12859-017-1537-8

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/ng.806

Druml, T., Baumung, R., & Sölkner, J. (2009). Pedigree analysis in the Austrian Noriker draught horse: Genetic diversity and the impact of breeding for coat colour on population structure. *Journal of Animal Breeding and Genetics*, *126*(5), 348–356. https://doi.org/10.1111/j.1439-0388.2008.00790.x

Druml, T., Neuditschko, M., Grilz-Seger, G., Horna, M., Ricard, A., Mesarič, M., … Brem, G. (2018). Population Networks Associated with Runs of Homozygosity Reveal New Insights into the Breeding History of the Haflinger Horse. *Journal of Heredity*, *109*(4), 384–392. https://doi.org/10.1093/jhered/esx114

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Francioli, L. C., Cretu-Stancu, M., Garimella, K. V, Fromer, M., Kloosterman, W. P., Wijmenga, C., … de Bakker, P. I. (2017). A framework for the detection of de novo mutations in family-based sequencing data. *European Journal of Human Genetics*, *25*(2), 227–233. https://doi.org/10.1038/ejhg.2016.147

Francioli, L. C., Menelaou, A., Pulit, S. L., Van Dijk, F., Palamara, P. F., Elbers, C. C., … Wijmenga, C. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825. https://doi.org/10.1038/ng.3021

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. 1–9. Retrieved from http://arxiv.org/abs/1207.3907

Gómez-Romero, L., Palacios-Flores, K., Reyes, J., García, D., Boege, M., Dávila, G., … Palacios, R. (2018). Precise detection of de novo single nucleotide variants in human genomes. *Proceedings of the National Academy of Sciences*, *115*(21), 5516–5521. https://doi.org/10.1073/pnas.1802244115

Grilz-Seger, G., Druml, T., Neuditschko, M., Dobretsberger, M., Horna, M., & Brem, G. (2019). High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. *BMC Genomics*, *20*(1), 1–17. https://doi.org/10.1186/s12864-019-5564-x

Grilz-Seger, G., Mesarič, M., Cotman, M., Neuditschko, M., Druml, T., & Brem, G. (2018). Runs of Homozygosity and Population History of Three Horse Breeds With Small Population Size. *Journal of Equine Veterinary Science*, *71*, 27–34. https://doi.org/10.1016/j.jevs.2018.09.004

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

Hadfield, J. (2016). Increased read duplication on patterned flowcells- understanding the impact of Exclusion Amplification. Retrieved from http://core-genomics.blogspot.com/2016/05/increased-read-duplication-on-patterned.html

Haldane, J. B. S. (1947). the Mutation Rate of the Gene for Haemophilia, and Its Segregation Ratios in Males and Females. *Annals of Eugenics*, *13*(1), 262–271. https://doi.org/10.1111/j.1469-1809.1946.tb02367.x

Jagannathan, V., Gerber, V., Rieder, S., Tetens, J., Thaller, G., Drögemüller, C., & Leeb, T. (2018). Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Animal Genetics*, (January 2019). https://doi.org/10.1111/age.12753

Jin, Z. B., Li, Z., Liu, Z., Jiang, Y., Cai, X. B., & Wu, J. (2017). Identification of de novo germline mutations and causal genes for sporadic diseases using trio-based whole-exome/genome sequencing. *Biological Reviews*. https://doi.org/10.1111/brv.12383

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., … Stefansson, K. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, *549*, 519. Retrieved from http://dx.doi.org/10.1038/nature24018

Kalbfleisch, T. S., Rice, E., DePriest, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., … MacLeod, J. N. (2018). EquCab3, an Updated Reference Genome for the Domestic Horse. *BioRxiv*, (April), 306928. https://doi.org/10.1101/306928

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., & Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PloS One*, *5*(11), e13996. https://doi.org/10.1371/journal.pone.0013996

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father-s age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/nature11396

Kuderna, L. F. K., Albrechtsen, A., Thèves, C., Hofreiter, M., Rieder, S., Librado, P., … Al-Rasheid, K. (2017). Ancient genomic changes associated with domestication of

the horse. *Science*, *356*(6336). https://doi.org/10.1126/science.aam5298

Lanata, A., Guidi, A., Valenza, G., Baragli, P., & Scilingo, E. P. (2016). Quantitative heartbeat coupling measures in human-horse interaction. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2016-Octob*, 2696–2699. https://doi.org/10.1109/EMBC.2016.7591286

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *00*(00), 1–3. Retrieved from http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*(20), 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T., & Hurles, M. (2016). Striking differences in patterns of germline mutation between mice and humans. *BioRxiv*, 082297. https://doi.org/10.1101/082297

Loewe, L., & Hill, W. G. (2010). The population genetics of mutations: Good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 365, pp. 1153–1167. https://doi.org/10.1098/rstb.2009.0317

Lorenzen, E. D., Nogués-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., … Willerslev, E. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, *479*(7373), 359–364. https://doi.org/10.1038/nature10574

Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., … Schierup, M. H. (2017). Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature*, *548*(7665), 87–91. https://doi.org/10.1038/nature23264

McCoy, A. M., & McCue, M. E. (2014). Validation of imputation between equine genotyping arrays. *Animal Genetics*, Vol. 45, p. 153.

https://doi.org/10.1111/age.12093

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., … Mickelson, J. R. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genetics*, *8*(1), 1002451. https://doi.org/10.1371/journal.pgen.1002451

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., … Wilson, J. F. (2008). Runs of Homozygosity in European Populations. *American Journal of Human Genetics*, *83*(3), 359–372. https://doi.org/10.1016/j.ajhg.2008.08.007

Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., … Waterston, R. H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. https://doi.org/10.1038/nature04072

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Narasimhan, V. M., Rahbari, R., Scally, A., Wuster, A., Mason, D., Xue, Y., … Durbin, R. (2017). Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00323-y

Nielsen, R., & Slatkin, M. (2013). *An Introduction to Population Genetics Theory and Applications*. Sinauer Associates.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., … Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*(7456), 74–78. https://doi.org/10.1038/nature12323

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178. https://doi.org/10.1101/201178

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., … Pääbo, S. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, *486*(7404), 527–531. https://doi.org/10.1038/nature11128

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-

based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., & Conrad, D. F. (2013). DeNovoGear: De novo indel and point mutation discovery and phasing. *Nature Methods*, *10*(10), 985–987. https://doi.org/10.1038/nmeth.2611

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., … Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Sandmann, S., De Graaf, A. O., Karimi, M., Van Der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, *7*, 1–12. https://doi.org/10.1038/srep43169

Schaefer, R. J., Schubert, M., Bailey, E., Bannasch, D. L., Barrey, E., Bar-Gal, G. K., … McCue, M. E. (2017). Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3943-8

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., … Orlando, L. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), E5661-9. https://doi.org/10.1073/pnas.1416991111

Shendure, J., & Akey, J. M. (2015). The origins, determinants, and consequences of human mutations. *Science*, Vol. 349, pp. 1478–1483. https://doi.org/10.1126/science.aaa9119

Smeds, L., Qvarnström, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, *26*(9), 1211–1218. https://doi.org/10.1101/gr.204669.116

Tatsumoto, S., Go, Y., Fukuta, K., Noguchi, H., Hayakawa, T., Tomonaga, M., … Fujiyama, A. (2017). Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-13919-7

Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., … Yagi, T. (2015). Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Research*,

*25*(8), 1125–1134. https://doi.org/10.1101/gr.186148.114

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., … Lindblad-Toh, K. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, *326*(5954), 865–867. https://doi.org/10.1126/science.1178158

Wei, Q., Zhan, X., Zhong, X., Liu, Y., Han, Y., Chen, W., & Li, B. (2015). A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics*, *31*(9), 1375–1381. https://doi.org/10.1093/bioinformatics/btu839

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, *16*, 15–24. https://doi.org/10.1016/j.csbj.2018.01.003

Zechner, P., Solkner, J., Bodo, I., Druml, T., Baumung, R., Achmann, R., … Brem, G. (2002). A nalysis of diversity and population structure in the Lipizzan horse breed based on pedigree information. *Livestock Production Science*, *77*, 137–146.

Zhang, C., Ni, P., Ahmad, H. I., Gemingguli, M., Baizilaitibei, A., Gulibaheti, D., … Zhao, S. (2018). Detecting the Population Structure and Scanning for Signatures of Selection in Horses ( *Equus caballus* ) From Whole-Genome Sequencing Data. *Evolutionary Bioinformatics*, *14*, 117693431877510. https://doi.org/10.1177/1176934318775106

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.25

Zook, J. M., McDaniel, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., … Consortium, G. in a B. (2018). Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. *BioRxiv*, 281006. https://doi.org/10.1101/281006

# APPENDICES

# APPENDIX A

# USED TOOLS

**Table A.A.1:** Programs and their versions that used in the analysis.

| Program | Version |
|---|---|
| **SAMtools** | 1.4 |
| **Bedtools** | 2.25.0 |
| **FastQC** | 0.11.5 |
| **MultiQC** | v1.7.dev0 |
| **Trimmomatic** | 0.36 |
| **Picard Tools** | 1.119 |
| **R** | 3.2 |
| **GATK** | 3.8-0 |
| **Freebayes** | v0.9.10-3-g47a713e |
| **Python** | 2.7, 3 |
| **Remap** | 2.1 |
| **IGV** | 2.4.10 |
| **Primer3** | v0.4.0 |
| **Codon Code Aligner** | 3.0.1 |
| **Bcftools** | 1.9 |
| **Vcftools** | 0.1.13 |
| **Plink** | 1.07 |

**Table A.A.2:** Usage purposes and command line usages of each tool.

| Tool/Algorithm | Process | Command line |
|---|---|---|
| **SAMtools** | Sorting | samtools sort @2 input.bam –o output.bam |
| **BEDtools** | Converting bam file to fastq file | bedtools bamtofastq –i input.bam –fq read1.fq –fq2 read.fq |
| **FastQC** | FatsQC analysis | fastqc –o . –f fastq input.fq.gz –t 5 |
| **Trimmomatic** | Trimming data | trimmomatic-0.36.jar PE -threads 6 read1.fq.gz read2.fq.gz read1_paired.fq.gz read1_unpaired.fq.gz read2_paired.fq.gz read2_unpaired.fq.gz TRAILING:10 MINLEN:50 SLIDINGWINDOW:5:20 |
| **MultiQC** | Reporting FastQC results | multiqc . |
| **bwa** | Indexing reference | bwa index -a bwtsw reference.fa |
| **bwa** | Mapping files to the reference | bwa mem -R '@RG\tID:laneID\tSM:horseID' -M -t 8 reference.fa read1.fq.gz read2.fq.gz > mappedFile.sam |
| **SAMtools** | Convert sam to bam | samtools view -b –S mappedFile.sam > mappedFile.bam |
| **SAMtools** | Merging different lanes | samtools merge merged.bam lane1.bam lane2.bam lane3.bam |
| **SAMtools** | Taking properly paired reads | samtools view –B –F 4 -f 2 –h input.bam > output.bam |
| **Picard Tools** | Removing duplicates | java -jar MarkDuplicates.jar I=inpıt.bam O=output.bam METRICS_FILE=metrics.txt REMOVE_DUPLICATES=TRUE |
| **SAMtools** | Filtering reads for quality | samtools view –q 20 –b input.bam > output.bam |
| **Picard Tools** | Calculating insert sizes | java -jar CollectInsertSizeMetrics.jar I=input.bam O=insert_size_metrics.txt H=insert_size_histogram.pdf M=0.5 |
| **SAMtools** | Indexing bam files | samtools index input.bam |

| | | |
|---|---|---|
| **Bedtools** | Calculating genomic coverages | bedtools genomecov –ibam input.bam –g reference.fa > output.txt |
| **Bedtools** | Calculating positional coverages | bedtools genomecov –ibam input.bam –g reference.fa –d > output.txt |
| **GATK HaploTypeCaller** | Calling variants | java -jar GenomeAnalysisTK.jar –T HaplotypeCaller<br><br>--(fix_misencoded_quality_scores) -R reference.fa -I son.bam -I father.bam -I mother.bam -o variants.vcf |
| **Platypus** | Calling variants | python Platypus.py callVariants --bamFiles=son.bam,father.bam,mother.bam --refFile=reference.fa --output=variants.vcf |
| **Freebayes** | Calling variants | freebayes –f reference.fa –L bamList.txt > variants.vcf |
| **Bedtools** | Intersecting bed files | Bedtools intersect –a file1.bed –b file2.bed > file3.bed |
| **GATK VariantAnnotator** | Annotating variants | java -jar GenomeAnalysisTK.jar –T VariantAnnotator<br><br>--(fix_misencoded_quality_scores) -R reference.fa -I trioBams.list –V variants.vcf –o annotatedvaariatns.vcf –A AlleleBalanceBySample |
| **GATK VariantFiltration** | Filtering variants in vcf | java -jar GenomeAnalysisTK.jar –T VariantFiltration -R reference.fa –V variants.vcf --filterExpression "QD < 2.0 \|\| MQ < 40.0 \|\| FS > 60.0 \|\| SOR > 3.0 \|\| MQRankSum < -12.5 \|\| ReadPosRankSum < -8.0 \|\| QUAL < 30 \|\| AC > 4 " --filterName "my_snp_filter –o output.vcf |
| **GATK SelectVariants** | Selecting heterozygous variants | java -jar GenomeAnalysisTK.jar –T SelectVariants -R reference.fa –V variants.vcf -select 'vc.getGenotype("horseID").isHet()' –o output.vcf |

| | | |
|---|---|---|
| **GATK VariantFiltration** | Filtering genotype of different idividuals in vcf | java -jar GenomeAnalysisTK.jar –T VariantFiltration -R reference.fa –V variants.vcf<br><br>--genotypeFilterExpression "AB < 0.3 \|\| AB > 0.75"<br><br>--genotypeFilterName "AB_filter"<br><br>--genotypeFilterExpression "DP > 34/24/27 \|\| DP < 10"<br><br>--genotypeFilterName "DP_filter"<br><br>--genotypeFilterExpression "GQ < 40.0"<br><br>--genotypeFilterName "GQ_filter"<br><br>-o output.vcf |
| **GATK PhaseByTransmission** | Calling de novo candidates | java -jar GenomeAnalysisTK.jar –T PhaseByTransmission -R reference.fa –V variants.vcf<br><br>--MendelianViolationsFile mendelianViolations -ped trio.ped -o output.vcf |
| **SAMtools, Bcftools** | Preparing consensus sequences for psmc | samtools mpileup -C50 –uf reference.fa input.bam \| bcftools call –c - \| vcfutils.pl vcf2fq –d 10 –D 34/24/27 \| gzip > output.fq.gz |
| **PSMC** | Converting fastq to psmcfa | fq2psmcfa -q20 file.fq.gz > file.psmcfa |
| **PSMC** | Psmc analysis | psmc -N25 -t15 -r5 -p "4+25*2+4+6" –o output.psmc input.psmcfa |
| **Vcftools** | Converting vcf top link format | vcftools --vcf input.vcf --out output --plink |
| **Plink** | Converting file to Plink's bed format | plink --noweb --horse --file input --out output --make-bed |
| **Plink** | Roh analysis | plink --noweb --horse --bfile inputFile --homozyg --homozyg-kb 500 --homozyg-snp 50 --homozyg-window-snp 50 --homozyg-window-het 3 --homozyg-window-missing 3 --homozyg-window-threshold 0.05 --homozyg-density 50 |

# APPENDIX B

## FILTERING DETAILS

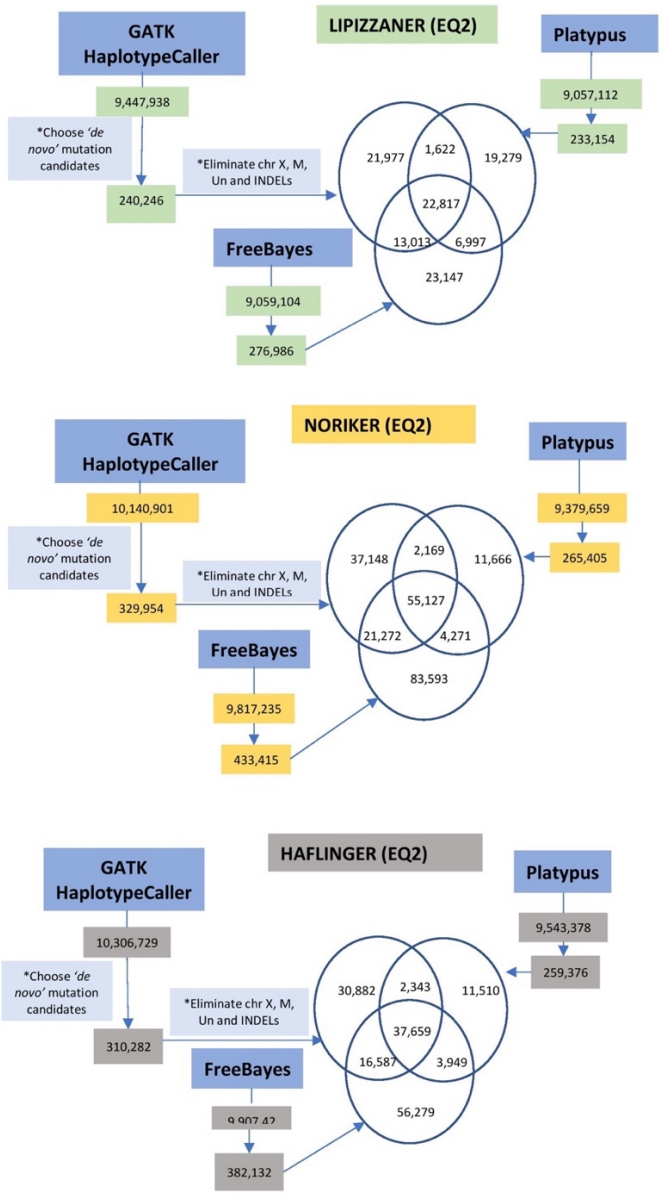


**Figure A.B.1:** Detailed results of first filtering step in Figure 4.4 for EQ2 mapped data.

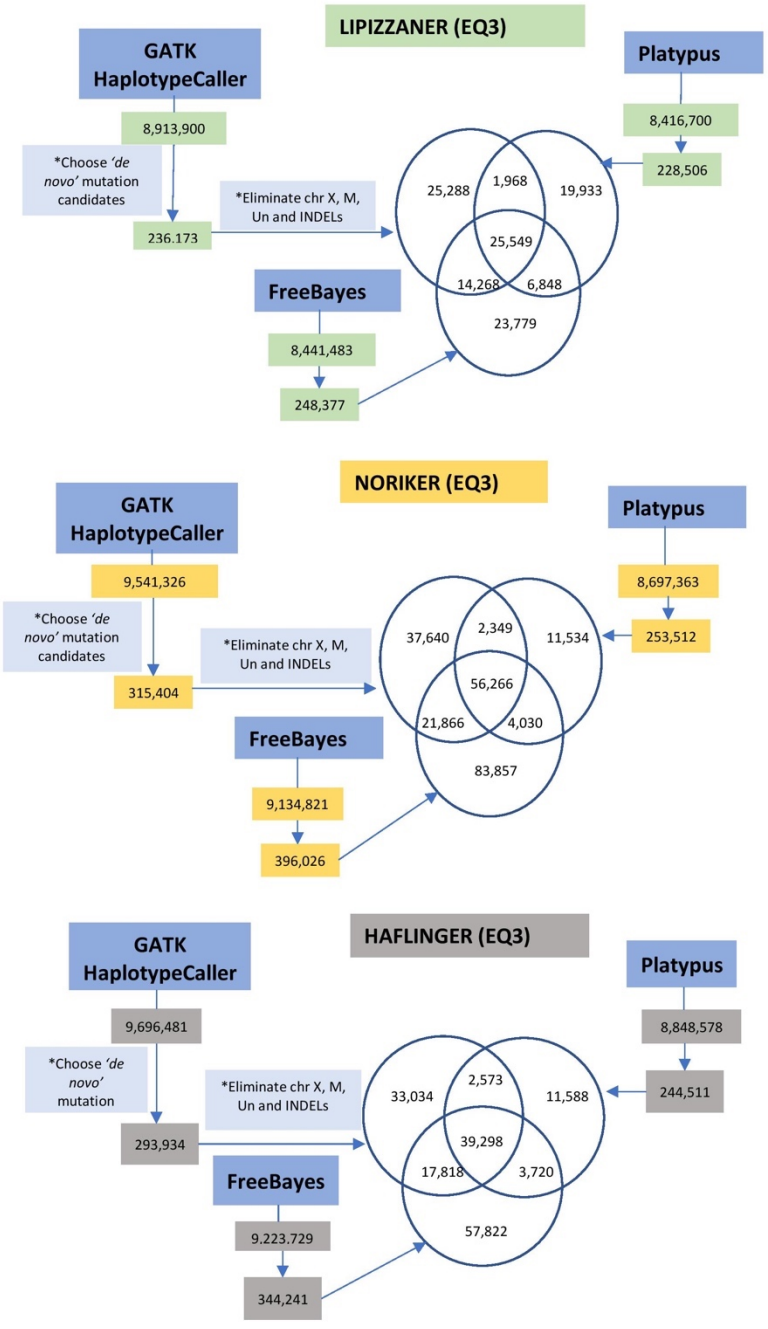


**Figure A.B.2:** Detailed results of first filtering step in Figure 4.4 for EQ3 mapped data.

# APPENDIX C

## DETAILED INFORMATION ABOUT DE NOVO CANDIDATES

**Table A.C.1:** Position and comparison information of all candidate 'de novo' mutations. Some of the reads mapped in one genome version (for example EquCab2) were mapped more then one position in the other genome verison (EquCab3). Such mutations (red labeled) were detected in the multiple positions when the position converted to the other reference version. The mutations in the fully red written rows cause inconsistency between comparison (EQ2, EQ3 and Jagannathan's panel) results. The choosen mutations for lab validation were indicated in the last coulmn. Green labeled candidates were choosen for future validation depending on their coexistence in both reference mapped data and not existence in Jagannathan data.

| TRIO | EQ2 chr | EQ2 pos | EQ3 chr | EQ3 pos | Ref | Alt | Found or not found in EQ3 | In Jagannathan data | Sanger validation done |
|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr1 | 15859530 | chr1 | 15978254 | C | G | found | no | yes |
| **LIP** | chr1 | 15864504 | chr1 | 15983228 | G | T | found | no | yes |
| **LIP** | chr1 | 68664504 | chr1 | 69214531 | A | C | found | no | yes |
| **LIP** | chr10 | 82141638 | chr10 | 83318763 | G | A | found | no | yes |
| **LIP** | chr13 | 27768579 | chr13 | 28928832 | T | G | found | no | yes |
| **LIP** | chr19 | 43987271 | chr19 | 46744475 | C | T | found | no | yes |
| **LIP** | chr2 | 49697936 | chr2 | 49938684 | G | T | found | no | yes |
| **LIP** | chr2 | 69806759 | chr2 | 70279783 | A | C | found | no | yes |
| **LIP** | chr4 | 96806258 | chr4 | 96948867 | C | T | found | no | yes |
| **LIP** | chr5 | 95223175 | chr5 | 92308816 | G | A | found | no | yes |
| **LIP** | chr8 | 29308865 | chr8 | 31918931 | G | A | found | no | yes |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr8 | 83490000 | chr8 | 86977321 | G | A | found | no | yes |
| **LIP** | chr9 | 40934406 | chr9 | 42581656 | T | A | found | no | yes |
| **LIP** | chr1 | 10431748 | chr1 | 10544851 | G | A | found | no | no |
| **LIP** | chr17 | 29582244 | chr17 | 29483913 | G | A | found | yes | yes |
| **LIP** | chr26 | 33187017 | chr26 | 34447304 | T | C | found | yes | yes |
| **LIP** | chr30 | 4510951 | chr30 | 4795900 | C | T | found | yes | yes |
| **LIP** | chr4 | 47113988 | chr4 | 47285204 | A | T | found | yes | yes |
| **LIP** | chr4 | 95252068 | chr4 | 95364574 | A | T | found | yes | yes |
| **LIP** | chr9 | 3085222 | chr9 | 3118093 | C | A | found | yes | yes |
| **LIP** | chr1 | 122283468 | chr1 | 123426208 | A | C | found | yes | no |
| **LIP** | chr14 | 52250726 | chr14 | 51556138 | C | T | found | yes | no |
| **LIP** | chr14 | 53324326 | chr14 | 52653121 | A | G | found | yes | no |
| **LIP** | chr17 | 16573635 | chr17 | 16642725 | T | C | found | yes | no |
| **LIP** | chr18 | 15530280 | chr18 | 15595862 | T | A | found | yes | no |
| **LIP** | chr2 | 4044657 | chr2 | 4059259 | A | C | found | yes | no |
| **LIP** | chr2 | 76393138 | chr2 | 76471715 | G | C | found | yes | no |
| **LIP** | chr2 | 107860807 | chr2 | 108203154 | A | C | found | yes | no |
| **LIP** | chr20 | 29358869 | chr20 | 30270730 | C | T | found | yes | no |
| **LIP** | chr22 | 573161 | chr22 | 619278 | A | G | found | yes | no |
| **LIP** | chr22 | 23971258 | chr22 | 24676750 | A | G | found | yes | no |
| **LIP** | chr22 | 23971258 | NW_019646087.1 | 8781 | A | G | not found | yes | yes |
| **LIP** | chr26 | 40048597 | chr26 | 41317337 | A | T | found | yes | no |
| **LIP** | chr3 | 81356650 | chr3 | 83166511 | G | A | found | yes | no |
| **LIP** | chr31 | 2013281 | chr31 | 8651746 | T | C | found | yes | no |
| **LIP** | chr4 | 9684807 | chr4 | 9719703 | C | G | found | yes | no |
| **LIP** | chr5 | 30402947 | chr5 | 27642878 | A | C | found | yes | no |
| **LIP** | chr7 | 44303415 | chr7 | 45561401 | A | C | found | yes | no |
| **LIP** | chr7 | 77372524 | chr7 | 79547823 | T | G | found | yes | no |
| **LIP** | chr14 | 296600 | NW_019643582.1 | 163072 | T | G | not found | no | yes |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr18 | 56233143 | chr18 | 56332502 | C | T | not found | no | yes |
| **LIP** | chr2 | 13626858 | chr2 | 13676181 | G | C | not found | no | yes |
| **LIP** | chr1 | 115504590 | NW_019646122.1 | 28415 | T | C | not found | no | no |
| **LIP** | chr1 | 115504590 | chr1 | 116526282 | T | C | not found | no | no |
| **LIP** | chr1 | 115504590 | chr1 | 116643212 | T | C | not found | no | no |
| **LIP** | chr15 | 18516837 | NW_019643418.1 | 116109 | T | G | not found | no | no |
| **LIP** | chr15 | 18516837 | NW_019643394.1 | 211146 | T | G | not found | no | no |
| **LIP** | chr15 | 18516837 | chr15 | 19197504 | T | G | not found | no | no |
| **LIP** | chr18 | 9517699 | chr18 | 9673823 | G | C | not found | no | no |
| **LIP** | chr2 | 86268 | chr2 | 80336 | A | C | not found | no | no |
| **LIP** | chr2 | 26636712 | chr2 | 26706669 | G | A | not found | yes | yes |
| **LIP** | chr21 | 1082137 | chr21 | 133290 | G | T | not found | yes | no |
| **LIP** | chr7 | 53003749 | chr7 | 55021516 | A | G | not found | yes | no |
| **LIP** | chr7 | 53003761 | chr7 | 55021528 | G | T | not found | yes | no |
| **LIP** | chr8 | 1656435 | chr8 | 3462742 | C | T | not found | yes | no |
| **LIP** | chr8 | 4013962 | chr8 | 6017687 | A | C | not found | yes | no |
| **LiP** | chr1 | 113086506 | chr1 | 114032630 | C | G | just in EquCab3 | no | no |
| **LiP** | chr1 | 163121631 | chr1 | 165088722 | C | T | just in EquCab3 | no | no |
| **LiP** | chr11 | 60729384 | chr11 | 61092938 | C | G | just in EquCab3 | yes | no |
| **LiP** | chr13 | 12281767 | chr13 | 12591780 | T | A | just in EquCab3 | yes | no |
| **LiP** | chr14 | 53281174 | chr14 | 52609933 | G | T | just in EquCab3 | yes | no |
| **LiP** | noalign | noalign | chr15 | 14765848 | C | T | just in EquCab3 | no | no |
| **LiP** | noalign | noalign | chr16 | 40321052 | C | G | just in EquCab3 | no | no |
| **LiP** | noalign | noalign | chr16 | 40321063 | C | T | just in EquCab3 | no | no |
| **LiP** | chr18 | 98589 | chr17 | 112172 | T | C | just in EquCab3 | yes | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LiP** | chr17 | 22409571 | chr17 | 22310438 | A | G | just in EquCab3 | no | no |
| **LiP** | chr2 | 63488928 | chr2 | 63948251 | C | T | just in EquCab3 | yes | no |
| **LiP** | chr20 | 30303426 | chr20 | 31215223 | A | G | just in EquCab3 | no | no |
| **LiP** | chr20 | 41448335 | chr20 | 42348483 | C | G | just in EquCab3 | no | no |
| **LiP** | scaffold_88 | 538547 | chr21 | 762973 | T | C | just in EquCab3 | yes | no |
| **LiP** | scaffold_190 | 742 | chr21 | 762973 | T | C | just in EquCab3 | yes | no |
| **LiP** | chr21 | 1345122 | chr21 | 1955710 | G | A | just in EquCab3 | yes | no |
| **LiP** | chr21 | 46651305 | chr21 | 47934770 | T | C | just in EquCab3 | no | no |
| **LiP** | scaffold_117 | 54314 | chr24 | 207208 | C | G | just in EquCab3 | yes | no |
| **LiP** | scaffold_117 | 54302 | chr24 | 207220 | G | C | just in EquCab3 | yes | no |
| **LiP** | chr24 | 3434129 | chr24 | 3299008 | T | C | just in EquCab3 | no | no |
| **LiP** | chr26 | 39072509 | chr26 | 40344282 | T | C | just in EquCab3 | yes | no |
| **LiP** | scaffold_180 | 61745 | chr29 | 552846 | C | T | just in EquCab3 | yes | no |
| **LiP** | chr3 | 85720365 | chr3 | 87844985 | G | A | just in EquCab3 | no | no |
| **LiP** | chr4 | 89013190 | chr4 | 89127688 | A | G | just in EquCab3 | yes | no |
| **LiP** | chr5 | 6347982 | chr5 | 6363111 | G | A | just in EquCab3 | no | no |
| **LiP** | chr5 | 49245871 | chr5 | 45822661 | G | C | just in EquCab3 | yes | no |
| **LiP** | noalign | noalign | chr6 | 28015944 | G | T | just in EquCab3 | yes | no |
| **LiP** | chr8 | 82614 | chr8 | 1810043 | A | T | just in EquCab3 | yes | no |
| **LiP** | scaffold_187 | 8272 | chr8 | 1897212 | G | A | just in EquCab3 | no | no |
| **LiP** | scaffold_103 | 55207 | chr8 | 1990358 | G | C | just in EquCab3 | yes | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LiP** | scaffold_107 | 177061 | chr8 | 1990358 | G | C | just in EquCab3 | yes | no |
| **LiP** | chr8 | 1768323 | chr8 | 3580793 | C | T | just in EquCab3 | yes | no |
| **LiP** | chr8 | 2072691 | chr8 | 3580793 | C | T | just in EquCab3 | yes | no |
| **LiP** | chr8 | 2177024 | chr8 | 4179490 | T | C | just in EquCab3 | yes | no |
| **NOR** | chr11 | 7649446 | chr11 | 7638666 | C | T | found | no | yes |
| **NOR** | chr17 | 20005831 | chr17 | 19909100 | C | T | found | no | yes |
| **NOR** | chr18 | 21015205 | chr18 | 21081929 | C | G | found | no | yes |
| **NOR** | chr8 | 31590451 | chr8 | 34654749 | C | G | found | no | yes |
| **NOR** | chr23 | 8657794 | chr23 | 8072201 | A | G | found | no | no |
| **NOR** | chr1 | 115297935 | chr1 | 116319664 | C | T | found | yes | yes |
| **NOR** | chr10 | 21323697 | chr10 | 21613706 | C | A | found | yes | yes |
| **NOR** | chr11 | 38548898 | chr11 | 38843410 | C | T | found | yes | yes |
| **NOR** | chr17 | 4565574 | chr17 | 4636020 | G | C | found | yes | yes |
| **NOR** | chr29 | 1231244 | chr29 | 2241954 | C | A | found | yes | yes |
| **NOR** | chr7 | 64671737 | chr7 | 66853582 | T | C | found | yes | yes |
| **NOR** | chr11 | 41767183 | chr11 | 42088312 | C | T | found | yes | no |
| **NOR** | chr16 | 2803472 | chr16 | 4274662 | G | C | found | yes | no |
| **NOR** | chr19 | 36357963 | chr19 | 38941787 | G | A | found | yes | no |
| **NOR** | chr19 | 36796471 | chr19 | 39380133 | A | T | found | yes | no |
| **NOR** | chr19 | 58455252 | chr19 | 61335678 | C | T | found | yes | no |
| **NOR** | chr20 | 1889151 | chr20 | 2144556 | C | T | found | yes | no |
| **NOR** | chr20 | 24456926 | chr20 | 25180104 | A | G | found | yes | no |
| **NOR** | chr21 | 1018747 | chr21 | 204090 | C | G | found | yes | no |
| **NOR** | chr21 | 2605508 | chr21 | 3228070 | A | T | found | yes | no |
| **NOR** | chr21 | 31429891 | chr21 | 32453825 | T | A | found | yes | no |
| **NOR** | chr22 | 43893158 | chr22 | 44860978 | C | T | found | yes | no |
| **NOR** | chr31 | 16751314 | chr31 | 16791438 | A | G | found | yes | no |
| **NOR** | chr8 | 3702372 | chr8 | 5726905 | G | T | found | yes | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **NOR** | chr8 | 4890346 | chr8 | 7004112 | A | T | found | yes | no |
| **NOR** | chr8 | 4890346 | chr8 | 5851711 | A | T | not found | no | no |
| **NOR** | chr8 | 11081101 | chr8 | 13296362 | G | A | found | yes | no |
| **NOR** | chr8 | 11081113 | chr8 | 13296374 | C | T | found | yes | no |
| **NOR** | chr23 | 199384 | NW_019644111.1 | 13075 | C | G | not found | no | no |
| **NOR** | chr23 | 252882 | NW_019644035.1 | 7983 | C | T | not found | no | no |
| **NOR** | chr14 | 473741 | NW_019643582.1 | 34819 | T | A | not found | yes | no |
| **NOR** | chr27 | 596466 | NW_019645822.1 | 25127 | T | G | not found | no | no |
| **NOR** | chr27 | 596466 | NW_019643525.1 | 46174 | T | G | not found | no | no |
| **NOR** | chr27 | 596466 | chr27 | 598908 | T | G | not found | no | no |
| **NOR** | chr8 | 1799905 | chr8 | 3612240 | C | T | not found | no | yes |
| **NOR** | chr8 | 2565803 | chr8 | 4567587 | T | G | not found | no | no |
| **NOR** | chr6 | 3270286 | chr6 | 2944071 | T | C | not found | no | yes |
| **NOR** | chr23 | 7864535 | NW_019643688.1 | 274190 | G | A | not found | yes | no |
| **NOR** | chr14 | 9869853 | chr14 | 9098133 | A | C | not found | no | no |
| **NOR** | chr12 | 12435069 | chr26 | 15321210 | C | T | not found | no | no |
| **NOR** | chr10 | 13100245 | chr10 | 13279495 | G | A | not found | yes | no |
| **NOR** | chr10 | 13100245 | chr10 | 13110832 | G | A | not found | yes | no |
| **NOR** | chr10 | 14126430 | NW_019644715.1 | 18830 | A | T | not found | yes | no |
| **NOR** | chr10 | 14126430 | chr10 | 14310057 | A | T | not found | yes | no |
| **NOR** | chr10 | 14151126 | chr10 | 14333736 | G | T | not found | yes | yes |
| **NOR** | chr12 | 17455667 | chr12 | 20972468 | C | T | not found | no | yes |
| **NOR** | chr12 | 17467671 | chr12 | 20984465 | C | G | not found | no | yes |
| **NOR** | chr28 | 20549421 | chr28 | 21592630 | C | T | not found | no | no |
| **NOR** | chr12 | 22821184 | chr12 | 26369633 | C | T | not found | no | no |
| **NOR** | chr10 | 23731364 | chr10 | 24166815 | T | C | not found | no | yes |
| **NOR** | chr11 | 25089146 | chr11 | 25163875 | C | A | not found | no | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **NOR** | chr20 | 32276595 | NW_019645272.1 | 15516 | T | C | not found | no | yes |
| **NOR** | chr20 | 32276595 | NW_019644819.1 | 52034 | T | C | not found | no | no |
| **NOR** | chr20 | 32276595 | chr20 | 33276510 | T | C | not found | no | no |
| **NOR** | chr20 | 32276595 | chr20 | 33139015 | T | C | not found | no | no |
| **NOR** | chr22 | 36838245 | chr22 | 37737690 | T | C | not found | yes | no |
| **NOR** | chr23 | 40878726 | chr23 | 40553043 | G | A | not found | no | yes |
| **NOR** | chr23 | 40878771 | chr23 | 40553088 | C | T | not found | no | no |
| **NOR** | chr11 | 41656509 | chr11 | 41981138 | T | A | not found | no | no |
| **NOR** | chr7 | 43560898 | chr7 | 44599012 | T | C | not found | yes | yes |
| **NOR** | chr1 | 44584336 | chr1 | 44944788 | G | A | not found | yes | no |
| **NOR** | chr1 | 44584336 | chr1 | 44913757 | G | A | not found | yes | no |
| **NOR** | chr8 | 49612506 | chr8 | 52686429 | G | T | not found | no | yes |
| **NOR** | chr16 | 65435532 | chr16 | 67018600 | T | C | not found | no | no |
| **NOR** | chr3 | 97439027 | chr3 | 99255299 | A | G | not found | no | no |
| **NOR** | chr1 | 155138900 | NW_019646251.1 | 75746 | G | C | not found | no | yes |
| **NOR** | chr1 | 155138900 | NW_019643074.1 | 75753 | G | C | not found | no | no |
| **NOR** | chr1 | 155138915 | NW_019646251.1 | 75761 | T | C | not found | no | no |
| **NOR** | chr1 | 155138915 | NW_019643074.1 | 75768 | T | C | not found | no | no |
| **NOR** | chr1 | 114280422 | chr1 | 115300895 | A | C | just in EquCab3 | yes | no |
| **NOR** | chr1 | 114234815 | chr1 | 115300895 | A | C | just in EquCab3 | yes | no |
| **NOR** | chr1 | 157856346 | chr1 | 159473532 | A | G | just in EquCab3 | yes | no |
| **NOR** | chr11 | 41636525 | chr11 | 41934046 | C | T | just in EquCab3 | yes | no |
| **NOR** | chr12 | 12410783 | chr12 | 12632000 | A | G | just in EquCab3 | yes | no |
| **NOR** | chr12 | 17467709 | chr12 | 20984503 | C | T | just in EquCab3 | no | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **NOR** | chr12 | 17368623 | chr12 | 20984503 | C | T | just in EquCab3 | no | no |
| **NOR** | chr12 | 17356600 | chr12 | 21028811 | C | T | just in EquCab3 | no | no |
| **NOR** | chr12 | 17356726 | chr12 | 21028937 | T | A | just in EquCab3 | no | no |
| **NOR** | chr12 | 32767000 | chr12 | 35488348 | C | T | just in EquCab3 | yes | no |
| **NOR** | chr13 | 8283509 | chr13 | 8571371 | A | G | just in EquCab3 | no | no |
| **NOR** | scaffold_294 | 47503 | chr13 | 9212502 | C | T | just in EquCab3 | yes | no |
| **NOR** | chr14 | 963402 | chr14 | 209691 | C | T | just in EquCab3 | yes | no |
| **NOR** | chr14 | 9078851 | chr14 | 8307528 | A | C | just in EquCab3 | no | no |
| **NOR** | chr14 | 92099865 | chr14 | 92754738 | T | A | just in EquCab3 | no | no |
| **NOR** | scaffold_340 | 31703 | chr15 | 35289 | T | C | just in EquCab3 | yes | no |
| **NOR** | chr18 | 71216313 | chr18 | 71333930 | C | T | just in EquCab3 | no | no |
| **NOR** | scaffold_131 | 79348 | chr21 | 173964 | G | A | just in EquCab3 | yes | no |
| **NOR** | chr24 | 294298 | chr24 | 153809 | T | A | just in EquCab3 | yes | no |
| **NOR** | chr24 | 38138711 | chr24 | 38509801 | G | A | just in EquCab3 | no | no |
| **NOR** | chr26 | 227587 | chr26 | 617551 | A | G | just in EquCab3 | no | no |
| **NOR** | scaffold_111 | 37009 | chr27 | 568230 | A | G | just in EquCab3 | yes | no |
| **NOR** | chr28 | 35964604 | chr28 | 37095887 | G | A | just in EquCab3 | yes | no |
| **NOR** | chr28 | 35990989 | chr28 | 37113509 | T | C | just in EquCab3 | yes | no |
| **NOR** | scaffold_93 | 491312 | chr29 | 891655 | A | G | just in EquCab3 | yes | no |
| **NOR** | chr14 | 565989 | chr29 | 1875093 | C | T | just in EquCab3 | no | no |
| **NOR** | chr29 | 770364 | chr29 | 1875093 | C | T | just in EquCab3 | no | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **HAF** | chr10 | 59172497 | chr10 | 60310985 | G | A | called | no | yes |
| **HAF** | chr17 | 1685198 | chr17 | 1753762 | T | C | called | no | yes |
| **HAF** | chr2 | 100761774 | chr2 | 101097784 | G | T | called | no | yes |
| **HAF** | chr3 | 53100714 | chr3 | 54576398 | C | A | called | no | yes |
| **HAF** | chr4 | 44983704 | chr4 | 45153777 | G | A | called | no | yes |
| **HAF** | chr5 | 21907765 | chr5 | 19148516 | A | G | called | no | yes |
| **HAF** | chr5 | 53320088 | chr5 | 49896738 | T | C | called | no | yes |
| **HAF** | chr7 | 91483698 | chr7 | 93696413 | T | G | called | no | yes |
| **HAF** | chr14 | 4822552 | chr14 | 4058364 | C | T | called | no | no |
| **HAF** | chr22 | 27863665 | chr22 | 28768316 | T | C | called | no | no |
| **HAF** | chr5 | 16057269 | chr19 | 3035926 | T | A | called | no | no |
| **HAF** | chr6 | 79058078 | chr6 | 80225422 | T | C | called | no | no |
| **HAF** | chr8 | 32485953 | chr8 | 35525782 | C | T | called | no | no |
| **HAF** | chr1 | 120334435 | chr1 | 121468596 | T | C | called | yes | yes |
| **HAF** | chr12 | 13784646 | chr12 | 14498413 | T | A | called | yes | yes |
| **HAF** | chr14 | 4827882 | chr14 | 4063623 | T | C | called | yes | yes |
| **HAF** | chr16 | 2815931 | chr16 | 4288452 | T | C | called | yes | yes |
| **HAF** | chr20 | 29311733 | chr20 | 30220575 | C | T | called | yes | yes |
| **HAF** | chr12 | 13825556 | chr12 | 14539782 | G | A | called | yes | no |
| **HAF** | chr13 | 19811302 | chr13 | 20890984 | A | G | called | yes | no |
| **HAF** | chr14 | 4827896 | chr14 | 4063637 | A | C | called | yes | no |
| **HAF** | chr17 | 39695905 | chr17 | 39599876 | T | C | called | yes | no |
| **HAF** | chr18 | 4152059 | chr18 | 4122138 | A | G | called | yes | no |
| **HAF** | chr20 | 26396785 | chr20 | 27301180 | A | T | called | yes | no |
| **HAF** | chr21 | 49759250 | chr21 | 51009814 | C | G | called | yes | no |
| **HAF** | chr22 | 552168 | chr22 | 598289 | G | A | called | yes | no |
| **HAF** | chr31 | 16997011 | chr31 | 17040610 | T | C | called | yes | no |
| **HAF** | chr6 | 13857585 | chr6 | 13645648 | T | C | called | yes | no |
| **HAF** | chr6 | 38444589 | chr6 | 39579282 | T | G | called | yes | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **HAF** | chr6 | 38444606 | chr6 | 39579299 | A | G | called | yes | no |
| **HAF** | chr7 | 9200806 | chr7 | 9677163 | A | G | called | yes | no |
| **HAF** | chr25 | 4135 | NW_019642813.1 | 331601 | G | A | not called | no | no |
| **HAF** | chr18 | 52005 | chr18 | 65931 | C | T | not called | no | yes |
| **HAF** | chr18 | 52005 | chr17 | 47646 | C | T | not called | no | no |
| **HAF** | chr24 | 540464 | chr24 | 405239 | A | G | not called | yes | yes |
| **HAF** | chr11 | 1039845 | chr11 | 1041777 | G | C | not called | no | no |
| **HAF** | chr19 | 4938297 | chr19 | 7325062 | C | A | not called | yes | no |
| **HAF** | chr10 | 7931916 | chr10 | 8047442 | G | A | not called | no | yes |
| **HAF** | chr23 | 8139725 | NW_019643688.1 | 553562 | C | T | not called | yes | no |
| **HAF** | chr18 | 9582002 | chr18 | 9033716 | A | G | not called | no | no |
| **HAF** | chr18 | 9582002 | chr18 | 12909306 | A | G | not called | no | no |
| **HAF** | chr12 | 14048779 | chr12 | 14953906 | C | T | not called | no | no |
| **HAF** | chr12 | 14048779 | chr12 | 14934992 | C | T | not called | no | no |
| **HAF** | chr12 | 14048779 | chr12 | 14879021 | C | T | not called | yes | no |
| **HAF** | chr12 | 15029981 | NW_019642075.1 | 5881 | T | C | not called | no | no |
| **HAF** | chr13 | 19315749 | chr13 | 20396513 | C | T | not called | yes | no |
| **HAF** | chr22 | 24976358 | chr22 | 25875039 | G | A | not called | no | no |
| **HAF** | chr22 | 24976376 | chr22 | 25875057 | C | T | not called | yes | yes |
| **HAF** | chr26 | 32427893 | chr26 | 33689194 | T | C | not called | no | no |
| **HAF** | chr26 | 32427893 | chr26 | 33752767 | T | C | not called | no | no |
| **HAF** | chr26 | 32427893 | chr26 | 33673583 | T | C | not called | no | no |
| **HAF** | chr7 | 34653199 | chr7 | 35818941 | A | G | not called | yes | no |
| **HAF** | chr7 | 34653210 | chr7 | 35818952 | A | G | not called | yes | no |
| **HAF** | chr2 | 42919845 | chr2 | 43141745 | G | T | not called | no | no |
| **HAF** | chr5 | 48921158 | chr5 | 45470619 | T | C | not called | no | no |
| **HAF** | chr3 | 54907512 | chr3 | 56379955 | C | T | not called | no | yes |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **HAF** | chr6 | 58762174 | NW_019642687.1 | 11347 | C | G | not called | no | no |
| **HAF** | chr7 | 59514418 | chr7 | 61694002 | T | C | not called | no | no |
| **HAF** | chr11 | 60756515 | chr11 | 61120069 | G | T | not called | no | no |
| **HAF** | chr18 | 65542248 | chr18 | 65657094 | A | G | not called | no | no |
| **HAF** | chr6 | 78694433 | chr6 | 79861978 | C | T | not called | yes | no |
| **HAF** | chr14 | 92829544 | chr14 | 93591845 | C | A | not called | no | no |
| **HAF** | chr14 | 92829544 | chr14 | 93486571 | C | A | not called | yes | no |
| **HAF** | chr14 | 92870205 | chr14 | 93560815 | G | A | not called | no | yes |
| **HAF** | chr1 | 143424677 | chr1 | 144891989 | C | A | not called | no | yes |
| **HAF** | chr1 | 162587371 | chr1 | 164688749 | T | C | not called | no | no |
| **HAF** | chr1 | 163046056 | chr1 | 165012407 | G | A | just in EquCab3 | no | no |
| **HAF** | chr11 | 52232188 | chr11 | 52582710 | A | G | just in EquCab3 | yes | no |
| **HAF** | chr11 | 52232228 | chr11 | 52582750 | C | T | just in EquCab3 | no | no |
| **HAF** | chr11 | 60729070 | chr11 | 61092624 | C | A | just in EquCab3 | yes | no |
| **HAF** | chr12 | 31057784 | chr12 | 34882799 | T | C | just in EquCab3 | yes | no |
| **HAF** | chr13 | 18159913 | chr13 | 19237651 | A | C | just in EquCab3 | yes | no |
| **HAF** | chr13 | 18160181 | chr13 | 19237678 | A | G | just in EquCab3 | yes | no |
| **HAF** | chr13 | 21927948 | chr13 | 23558006 | A | C | just in EquCab3 | yes | no |
| **HAF** | chr13 | 26958570 | chr13 | 26404373 | G | A | just in EquCab3 | yes | no |
| **HAF** | chr18 | 13036153 | chr18 | 12929475 | G | A | just in EquCab3 | yes | no |
| **HAF** | chr2 | 34726954 | chr2 | 34906791 | G | C | just in EquCab3 | no | no |
| **HAF** | chr20 | 31048188 | chr20 | 31912373 | G | T | just in EquCab3 | yes | no |
| **HAF** | scaffold_224 | 5223 | chr20 | 51466011 | T | G | just in EquCab3 | yes | no |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **HAF** | scaffold_88 | 561709 | chr21 | 788265 | G | A | just in EquCab3 | yes | no |
| **HAF** | scaffold_111 | 31059 | chr24 | 51881 | T | A | just in EquCab3 | no | no |
| **HAF** | chr14 | 673521 | chr24 | 291480 | A | T | just in EquCab3 | no | no |
| **HAF** | chr24 | 455123 | chr24 | 314661 | A | C | just in EquCab3 | no | no |
| **HAF** | scaffold_85 | 1123414 | chr27 | 26964 | C | T | just in EquCab3 | yes | no |
| **HAF** | scaffold_180 | 27440 | chr27 | 26964 | C | T | just in EquCab3 | yes | no |
| **HAF** | scaffold_196 | 74869 | chr27 | 79930 | C | A | just in EquCab3 | yes | no |
| **HAF** | noalign | noalign | chr28 | 18454 | A | G | just in EquCab3 | yes | no |
| **HAF** | chr6 | 38004787 | chr6 | 39063395 | T | A | just in EquCab3 | yes | no |
| **HAF** | chr6 | 38004798 | chr6 | 39063406 | A | G | just in EquCab3 | yes | no |
| **HAF** | scaffold_83 | 413845 | chr6 | 86814340 | A | G | just in EquCab3 | yes | no |
| **HAF** | chr8 | 70163030 | chr8 | 73645663 | A | C | just in EquCab3 | no | no |
| **HAF** | chr9 | 38880815 | chr9 | 40536852 | C | T | just in EquCab3 | yes | no |

**Table A.C.2:** Detailed information of 'de novo' candidates obtained from GATK's PhaseByTransmission (PBT) algorithm in Lipizzaner trio. 111, 113 and 166 were the sample IDs of Lipizzaner son, father, and mother respectively.

| TRIO | EQ2 chr | EQ2 pos | EQ3 chr | EQ3 pos | Ref | Alt | Depth (111/ 113/ 166) | Found in EQ2 finalized candidate list of Lip | Found in EQ2 finalized candidate list of Lip | Found in Jagannathan's variations |
|---|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr12 | 14923621 | chr26 | 15298236 | T | C | 75/135/142 | not found | not found | not found |
| **LIP** | chr12 | 30138931 | chr12 | 33942154 | C | A | 42/37/31 | not found | not found | found |
| **LIP** | chr26 | 14842365 | chr26 | 15341895 | A | C | 45/83/73 | not found | not found | found |
| **LIP** | chr26 | 14842365 | chr12 | 12626648 | A | C | 45/83/73 | not found | not found | not found |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr31 | 16520917 | chr31 | 16561393 | T | C | 453/368/326 | not found | not found | found |
| **LIP** | chr7 | 98533779 | NW_019643487.1 | 4355 | G | A | 29/62/93 | not found | not found | not found |

**Table A.C.3:** Detailed information of variants which were validated in the laboratory. Primer sequences and sequences produced by Sanger resequencing were seen.

| TRIO | EQ2 chr | EQ2 pos | Sanger result | Category | Variant in brakcets [] including flanking region | Primer forward ID | Primer forward seq. | Primer reverse ID | Primer reverse seq. | amplicon length (bp) |
|---|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr1 | 15859530 | SNP validated | True_denovo variant | atacctgtttcccctattactaacatcttacattagtatggtacatttgttacaattaatgacccagtattgatccattattagtaactaaagtccatagcttattcagattcccttagtttttacctaatgtccctttctgttttgggatcccatccaggacaccacattagatttagttgtgatgtctccttaggctcctcttggctgtgacagtttccaagacttcccttgtttttgatgaccttgacagttttgaggagtcctggtattttgaggactatattgtaggatgacc[c/g]tctcttggaatttgtctgatgtttttctcaggatttgactgaggttatgggttttggaagatcacagaggcaaagcaccatttgcatcacatcgtatcaaaggcacctgctatcagcttgatctatgattgttgatgttgatctttatcgcttggctgaagtggtgttggtcaagtttctccactgcagagttacttttctcctcctcctttccatactgtactccttggaaagaaatctacacatagcctactcctaaggagtaggggtgggggttatgctccatctcattgaggg | chr1:15859420fwd | CTCCTTAGGCTCCTCTTGGC | chr1:15859705rev | TTGACCAACACCACTTCAGC | 286 |
| **LIP** | chr1 | 15864504 | SNP validated | True_denovo variant | atagttaatgatgttgaacatctttcatgtgcctgttggccatctgtatattttctttggagaaaatatctgttcaggtcttttgcccagttttttaatgggttgttggtttttctgttgttgatttgtatgagttctttgtatattttgactcttaacccccttatcaaatatatggtttgcaaatatcttctctctattgttaggctgtctttttgttttgttgatggtttcctttgctgtgtagaagctttttactttggtgtagaccaatttgtttatttttttctattgtttccctt[g/t]cctggtcagacagggtacttgaaaaatatgctgctaagactaatgtcaaagagcgtactgcctatgtttttcttctagaagttttatggtttcaggtcttacattcaaatctttaatccattttgagttaatttttgtgtatggtgtaagataatagtccactttcattcttttgcatgtggctgtccagttttcccaacaccatgtattgaagagactttcctttctccattgtaccttcttaactctcttgtcaaaaattagctgtccatgtgctggccccgtagccaagtgcttaagtt | chr1:15864504.f | aatatctgttcaggtcttttgcc | chr1:15864504.r | ctacgggggccagcacatg | 524 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr1 | 68664504 | SNP validated | True_denovo variant | tgaagattggctacaaatgttagctcaggtgtcaatcttaaaaaaaaataaaattagaattaagtttttattcaaatatggtttattagagataatttatctggcatcaaactattgcataaagcatacgtgaggtaatcttttaatgataaatacaaacacagtcagcccagatcttcctagactttgtaatacttttgtaatactagcattaaagatagtaatgaagtgttcttgaaaaagctgttccataaccaaatatttgggaaacactgagtaaacaaagtaaaatacatta[a/c]aaaaatgcaagactccctctgagattcaaatattaatttgttaatatttatttgtattagttatctattgctatgtaacaaattgttttaaaatagtggctaaaaacaacaaggaatagtgaatgctgcttacagtttctgtgggtcaggaatttgtgagcagcttggttgaggggctctggctcagggtctctcacagagggacagtccagctgttggctggggaagccccatctgaaggcttgactggggctggaggaaccacttccaagatgcctcactcacttggttggcaagttg | chr1:68664364fwd | ACAGTCAGCCCAGATCTTCC | chr1:68664677rev | CTCAACCAAGCTGCTCACAA | 314 |
| **LIP** | chr10 | 82141638 | SNP validated | True_denovo variant | acaattctaagataaaggcattcctatgcaaaccgttaattgtaaaatgaaagacagaaatagcaaaagtggtcacaggcacaatagtgtacattacagcatattcttggctttatacatacttgaaattttccataatttaaaatgagttaaacttcaaatgaaaaagtacatttttctgaatactgaggatcacgtgccctgttaagagtcttatttttttaaacagagtaagcagctctaaatggaaacaaaacccaaataagcattaatcttaactagttggctcaaatagttac[g/a]tctgtggaacaagaaaaaaaatttccaactgcatcacgttggtgattaaaaagcatttttatcgctccagaaggtacaggtgtttctcaattctaaaatggaaacagcacgtgcccagactgctgatatgcaaaagtttgtcgtatcaaggaaacacaggatttatgagtcttgactcacataagtctcaaacgtcttgtgcttgttcagaggcaaatagaaggtcccaataaaaagagttcaggagaagaaagaagtcacacacttcctccctagcaattcacatcaagatagggaat | chr10:82141528fwd | AGGATCACGTGCCCTGTTAA | chr10:82141853rev | TGCCTCTGAACAAGCACAAG | 326 |
| **LIP** | chr13 | 27768579 | SNP validated | True_denovo variant | agaagacatttggactctctggggctcattttcctcatctgtaaaatggggctatgatagttcctacctcactgggctactgaaaggattaaatgggttaatccacataaagcatacagaaagacccgcctctagaaaacaactgctcaatgtggcttcattatcatcaccaccatcatcatccttgttgtaactctgtgcacagatgttatattttcttaatgtaaattcctaaaaataaaatgactaggccaaaagatatgatttttgatatatgacagtctgatcagcaatggagat[t/g]tattataattatgacaagttggtaagttgttttctcctgcggggaccccccgtgagtggtctgtggtctggagacacattttaagatcaggttcatccacgtcgtccccgcttctgtttcattaggcacagatgctcctgaacggtcccgagcctg | chr13:27768420fwd | ACTGCTCAATGTGGCTTCAT | chr13:27768809rev | CATTGCTCCTTCCGTTCCTT | 390 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ttttactctcctacgttatattcac caccctagacttcttaaagatcc cgcaggaaggaacggaaggga gcaatgcaggtggcggtggat gtggaggttctgctcccggcgt gtcctggacagagcgacctcc gaggaaagcaca | | | | | |
| **LIP** | chr19 | 43987271 | SNP validated | True_denovo variant | ggttattcctctgtgaatgcgact taggctgctttttttcctccctagc cagtctgttcacccttcattatctt tgttctcttcctatttttcacagtttc ttttttaaagattggcacctgagc taacaactgttgctaatctttttttt ttttttagctttttctccccaatccc cccagaacatagttgtatatttta gttgtgggtccttctagttgtggg atgtgggacaccgcctcaacat ggcctaacgagcagtgccatgt ccgcacccaggatccgaaccc tgggccgccaaag[c/t]ggaa tgcgcagacttagccactcggc cacagggccagcccctgttttc agtttctttccttcttttttatcttccc taaatgcccttctctctgttttttct ctttctctgatttttttgtcctcagtc tccttaccagtggttctggaact attgagtattccagaccagctga aggggttgatttgtgctggtcag gcctgagggcaggaagatcca tcctgtttggaggcctggaccct ctgtggcccaagccatttgcttt ggctttttcctagtaatgagatgg gagggctgcctcc | chr19:43987179fwd | CTTCTAGTTGTGGGATGTGGG | chr19:43987567rev | GCAGCCCTCCCATCTCATTA | 389 |
| **LIP** | chr2 | 49697936 | no SNP | | gacaaaccttttagctagagtca ccaagaaaaaaagagagaag gctcaaataaataaaatcaaaa atgaaagaggagagattacaa gggacacctcagaaatgcaaa agatgataagagaatactatga aaagctatacggcaacaaattg gataatatagaagaaatggata aattcttagaaacatacaaccttc caaaactggaccaagaagatgt agaaaatttgaatagaccaatca ccagtaaggagatcaaaacag caatcaaaaacctcccaaaaaa caaaagtccagggcca[g/t]a tggcttccctggtgaattctacc aaacattcaaagaagacttaata cctatccttctcaaactcttccaa aagattgaagaggaggggag gcttcctaactcattctacgaag ccaagatcattctgataccaaaa ccagacaaggacaacacaaaa aaaacaaaattacaggccaata tcactgaacatcgatgcaaaaat cctcaacaaaacactagcaaat cgaatacgataatacattaaaaa gatcatacatcatgatcaagcg ggtttcattccagggatgcagg gatggttc | chr2:49697936.f | aggagagattacaaggacacc | chr2:49697936.r | cctggaatgaaacccgcttg | 516 |
| **LIP** | chr2 | 69806759 | SNP validated | True_denovo variant | accatctgatgacttctcttttctc tgtgatgcaagagacaagccat ctgccaaaaatgaagataagg gtctgaaatttaaggtgagtaga aaatgttagaaacagttattgtg gaatgcagtgcagcaagctga ccacggaaatgtagcaggtttg ctggacagtgaagggctaactg agattggaggtaataaatttaaa agtggttccaatctgacttcttgc gtgttggccaggaacaataatg | chr2:69806611fwd | GTTTGCTGGACAGTGAAGGG | chr2:69806944rev | ATGTCGGACACCTACCACAG | 334 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | tgcctggatgaagacttcagga ttcttccagaacaatgagatgaa gacagag[a/c]agcaaggaa atctgtaatgttgacagcaattat ggaaataaagaattagagaatc tggggccagccgggtggtgca gcggttatgtgcgcatgttccgc atcagtggcctggggttcgccg gtttggatcccgggttgggacg tagcaccgcttggcacgccatg ctgtggtaggtgtccgacatata aagtagaggaagatgggcacg gatgttagctcagggccagtctt cctcagcaaaaagaggaagatt ggcagcagttagtttagggcta atcatcctaaaacaaaaaaaaa ag | | | | | |
| **LIP** | chr4 | 96806258 | SNP validated | True_denovo variant | tatgaaaacagaacgctgccac ttataagctgggacaaagacct cagacaaagaccttgcctacgg gactgtggtgaggactgatgag acaactcaatcttcaaagataca tgaacctaagctactagtatgtat tacagtactttcatatgagcagct cctaatagttctagaccttgta ggctatactatgtgtaagggtaa tgagctactgatcttagattaatg ttttccccaggctcttctctgtctt cgtataaaaagcagcaagtggct tggatctgcttttgcctaggaaa g[c/t]gactgcagtgtcctcctc aggcatatagacacaggtgtgc agttagttcttgcccttgaggac agcacgtatgagaggatcttttg taggattttagactgtcccttccc ttggcaggcagagcaattccctt gagctccagagccttcaacag gaatttgaagttagcgaacctgc ttgggaaggccaggagcctttg accaagatttgtagggaagaaa gttcactgaaaacctcaaaggg gatgcccccgtgaggaaggca caaatctgagctagggacctgg gttcaatataaagg | chr4:96806023fwd | GGACTGTGGTGAGGACTGAT | chr4:96806387rev | GAATTGCTCTGCCTGCCAAG | 365 |
| **LIP** | chr5 | 95223175 | SNP validated | True_denovo variant | acagcatgggagaaaatatttg cacaccatatgtctgataagga gttaatatccaaaatatataaag aactcatacaactcaatagcca aagaacaaataatccaattaaa aatgggcaaaggatctgaaca gacatttttccaaagaaaacata caaatggccaacagggacttg caaaggtgctcaacatcactaat caccacagaaatgcaaatcaaa accacagtgcgatatcacctcc cacctgtcagaacggcttatca aaaagacaagaggtagcaaat gctggtaaggatgc[g/a]gag aaaagggaacccatatacactg ttagtgagaatgtaaaatggtgt ggccattatggaaaacagtatg gcagtttcttaaaaaattaaaaat tgaactaccatgtgatccagca atcccactcctgggaatacatca gaaggaaaagaaatcagtatct cgaaggggtatctatacaccca tgttcactgcagcattattcacaa tagctaataaatggaaacaacct aggtgtctattgaaaagatgaat ggataaagatgtggtggaatatt attcagccaggagaaagaaag aaa | chr5:95223058fwd | GCTCAACATCACTAATCACCACA | chr5:95223302rev | AGGAGTGGGATTGCTGGATC | 245 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr8 | 29308865 | SNP validated | True_denovo variant | tctccgtgtgctctcgggcccgtccacgtggacatggagtggacgtgtctgagcttcctcaccaggtggggcttcaggacagcggacggctcacgtggcagttcacagggcagggccagcgccgggctggcgcagggcagaggtccacagcttcacctctgccctcctccatcgatgacccgccacaaaagcccagcccgctgccaggggagggacacaggctccacctcctgatgggggaaggttctagaagagatcactgcctccgtctttggaagcacaatctgcgccctggagtac[g/a]tgaggaaggtgagattctggaaggaaacgtgtgctgcagagcagggttcagcagacgagggccacagaccggatcctgcctcctgttctgcagatgagtgcagtcgggccacagccacacccgtccacgcacacgctgccctgtggctgctcccaacgtggcccggaggctacagtgcctactgtctggcctgagcaatggggagaaatacacactggtcgctgcccaggttcccgacacggagcctgagggccctcgtaatctcccgagggaccggcacggcaggagctcacacgcgg | chr8:29308598fwd | ATGGAGTGGACGTGTCTGAG | chr8:29308957rev | GCAGAAACAGGAGGCAGGAT | 360 |
| **LIP** | chr8 | 83490000 | SNP validated | True_denovo variant | tggcatgagatttcatcacaggacttagaaggatgctcaatttaaaacttatacattcttatttctggaatgttccatttaatttttttcagacctaagttgaccacgggtaactgaaaccgcagaaagcaaggggggctctactgtatgtagagtactctcagaactgctcactgctctattctgtgtagggattcggaagcccaggttccacttacagttctgcacaaaacagtccccatgcggttcagcctatcattagttaaagcagcagaaaggacgagatctctaagtcaattcca[g/a]ttcattttgagactttatctgcaaagaagtatgagaacttctttttccatctcttccaaatttttgccctatttgtgactttgtattgaattgaataggaaagaaaaccatgaaggatttttgtcaaagagggatacaatggccttctttgctttgatgttgggtgctgatcagggaagaaatatcacccagcatctgcaaccactgactgatggaaatcagttctttgattcctctcttattgtctccatgagtcactgcccagatgagagtctcatgcaaaaattagtcacagagtaa | chr8:83489881fwd | CTGTGTAGGGATTCGGAAGC | chr8:83490270rev | CTCATCTGGGCAGTGACTCA | 390 |
| **LIP** | chr9 | 40934406 | SNP validated | True_denovo variant | tatatagccatgccaatcaatcaataagttttaagatatctcaaaagaaaattagttagaaaattatgtcaaaaaaggttttatcataatagagattaacatttttttaagaatatgtattctccttcaggatattcccaatgaaagtaatctggccaaaattaaaattatcaaattgcaagattgaaaataaatttttagagactattatattccggcagatcatgcctcaagcgaattgctttcagcttctccacttcaatctttgctcttagcagctcttcttctagttgctgccttct[t/a]ttcaatccatccctcttctcctgaacatataatcccaaatttttttgattttctaaaattatctgatgttcctctttcagcatttgcagaatctgaggatcatacccctaattgtgacctaaagcgttctccactctcatgctgaaaaaaa | chr9:40934241fwd | CCCAATGAAAGTAATCTGGCCA | chr9:40934622rev | GCCGTCACTTTCCTCTGTTG | 382 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | tcatctctccgtgggatagaaga tggagacgatgctattctcatat caacagaggaaagtgacggc gacaaagatctttccgtaacgtg cactaattcatgttcttctctttgct ctaaagtatcactttgttgagaat ttaaaaccagt | | | | | |
| **LIP** | chr17 | 29582244 | SNP also in parent | ccttattcctgattttataggttgtt gcatcctccgttaggagcagct cattgtgatattgaagaattgctg caagaagatgacacaaacgca caagggtttaggtacacaactg gataactcaataaatgtggttga atgttacagtaatggcgctgctg aatcatatttccctctatctacac cctctattatttctcttccacattg actctggccttggctgtgtgactt gctttgggaacagggcattaac aaatgtgacccaagcagaggc atgacaaatactttctcatcagg[g/a]ccccatctcttgctgcttttg gaatacagccactatgtgaaga agtctcctgacctgctggcaac agacggcccagctgacagcca gcaacattcatcagacgtgtga atgatgcttagactaaccagcc ccattgcagctgatggataacta tagaaacatgagtgaccccag gtcagactagaaaaggaacctc ccagctgagctcagcccaaatt gctgacccacagaatcagagc aaataaaatgatagatatttaag ccattaaatattggattgatatgt agcaaaagatactt | chr17:29582244.f | ggttgttgca tcctccgtta | chr17:29582244.r | tgctctgat tctgtggg tca | 524 |
| **LIP** | chr26 | 33187017 | SNP also in parent | aaatacacatctacgattgttatg tcttctcaatgaagtcacccctttt atcattatctttatcttttctaactct ttgtcttgaagtctgttttattggat gttaataaagtcattccaaccttc ttaagcttactgtttgcatggtata tctttttccatccttttacttttaagc tatctatgtctttttctttttttaaag gaggtagtcctttgtttgaaatttt ttttcttcctaaagccccagtatat agttgtataacctagttacaagt ggttctagttcttctatgtggga[t/c]gccaccacagcagggcttg atgcacacagtacataggtctgt gcccaggatccgaaccagcaa accccgggctgccaaagcaga gcatgagaacttaacccctcag ccacagggccggccccctatct gtgtctttttaattaaagtgtgtcc taaagagaacatacactcaggt cttgcttttctgtttaatccattctg acattctctgccttttaactgcag ggcttagtccattatcatttaatgt aattactgataaggctgaattag gtctaccattctgttcatttttggtt t | chr26:33187017.f | cttaagctta ctgtttgcatg gt | chr26:33187017.r | tgataatg gactaagc cctgca | 421 |
| **LIP** | chr30 | 4510951 | SNP also in parent | cctttgaccctggacatctgacc ccagccaactgccaccaaggc tactccaacctccagtgtgctag gattctagaacaagcccagctc caactgtcactctgagtgtggcc tggcctggcctccctgagcctg tttcctcgtctggaaagggtgag tcctccctcctcacagggtct ctgaggactcctgccatgaccg tcatcactgccctcccctcaccc ctccaggggggagcagggcag aaagctccgcatgctcctctccc | chr30:4510951.f | ttgaccctgg acatctgacc | chr30:4510951.r | ggatgggt gttaggtg agca | 445 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | cgcttttagcccatcacccccct tgaggcctg[c/t]gccacagg accatcatgcctgcatctccccg atgaggagacagaggcccaac aggggcagcgagctgctcagg gccacagaggagaggccactt gcccctcccagccagaggccc tgttctcatgtgctcacctaacac ccatccccacccctgaccgga gccctgtgccctcagttctcagg tgtgtccttggccgagacgtgg atggagagggcggggtgtcct gggagcctgggtgagtggctc ctgcccggcccaggccctgga gagtctggcccccaggctgga acggaggccac | | | | | |
| **LIP** | chr4 | 47113988 | SNP also in parent | atatgaagcccaaaaaccaatc ctagggaacaaaatcaagtaca gtgaacagaaagaattctaaca ggtgcacatggaaagggcata gatgtaatgaatgttttgatgatc ttcccagtctccctccttgaaga ctaaaacacttattcctctagctg ttggaaatattgatgactcacaa cagctgggcagggaagagcct ctttgccctaggtcacagtttctc ccagggagcagctcacatccat taactgatcaattgaagataaaa gacttggctttctttgcttcaaca attac[a/t]ctcatgggccttct ggttctagagctccccatggga tcagctgtggccttttttcgagatt tcattaaaaagcaactcttctcaa tccaatttccttcactccctcact ggtttgataccaataaacttcctg cctttaactctccatctcagaatc ttttttccagggaacttgacctga aacaatatgctttgggagccca gagctggggttgaatccaagca atttaacttgctatgtgaccttgg gttttttagtctcctcagcctctgt cagcaaaaaatagtaatatgag ggcaa | chr4:47113988.f | cccagtctcc ctccttgaag | chr4:47113988.r | acccaag gtcacata gcaagt | 440 |
| **LIP** | chr4 | 95252068 | SNP also in parent | ttttactcagcatatggctcttga gatgcatccaagttgttacatgt atcaatagctcatttttttttctttttat tgctgagtagcattccatgacat ggatgtgccacagtttccttagc cattcgcctattataggacatttt ggttgtttctgtttctttttttacaat tacaagtaaagcagctatacatt aacatttacagaactctgtgtgg atgtaagtgctcgtttctctggga taaatttatgcccaggagtgaga ttgttgggctacaacataaatata ttcttatttttt[a/t]aagacagtg ccaaatttatttttctggagtggtt gcacctctttatattcccagagat tcatttttaataccagtgatttcaa ctacatgtaacattccacgttgtc atgtccataacacgcatgcgaa tgactaaaagaatagaaacctg gtagactcagagagtttttctttg gatgcttttgaactcttgtcttcag aaatatttgaattctgtctaacctt tatctaactggaggtctgcctctt caatgttaccatcactggctgtt gctgtctgcatgcttgtgtcagc cgtcct | chr4:95252068.f | tggctcttga gatgcatcca | chr4:95252068.r | tgacacaa gcatgca gacag | 581 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LIP** | chr9 | 3085222 | SNP also in parent | ttgatcttctgaaattgttttatgacattattaagaaatcgccatggaccagttaattagctcaactagttctttcacaaacatacgctaaagctattcaacatcaaaattctctgcattatttaaatatgttgaagtcacttccagtatttgcccagaagtcatgaaattatatatgtatatttgcatgtctgtgtatgtctgtatacatacatgtatgtatatatgaatatgtgtgtgttatgcatgaagatatttacgtgtgtgcacatatatatatataacttatacatacgtat[c/a]aaaacataaatacatatacataaaaagtttaaaataactttacagaaacaaaaaaagtcaagtccaatgcttgtttagggcaaaactgaatccagagtcccaagccttagcacattcacactctccagctagaaagatctcatgtcgaaacaatgaagggtccacagcacttgtgctgaattaaggcaaaggaggcaggaactctttctactgttggatgcattgcatgccttctttatcagcaagccttttcaaacaactgtcccttagctcattattcagatttgccacctaggag | chr9:3085222.f | gccatggaccagttaattagct | chr9:3085222.r | gctggagagtgtgaatgtgc | 393 |
| **LIP** | chr7 | 77372524 | bad sanger seq | cctttaatcagcccatttttgcttgaaatttatattatttctaatattctaccttcgcacaatgcttcaatgaacatcacttacggactctttacttactggtttgaatatttctgttggagaaatttatgtcaccatcacgtcaaaagacatgtctatgtttaacttgatagattctgtaacgttagtctctcaaaatcctgaacagtttcattccatcgatagtttcattgacaacataagcaataaagatatgatcagtatccgatgaagctaaattacatgcaatgtaatatc[t/g]atatggcctatgcattgatttccatatctcactaataataataataaataattagtatatgatctctacaagaaactttattagtactttatacatgttttctcatttaatcttcacatctggctatgtttgcacaacttttatggctgagaaaaagcagaatatttaattatctgaaggctatacagttagtaagtgatagagcctgaatttgaatccaagaagctgactttacagctgatagctgtttcttgtgtctgaatttcttatggatcaaaataatttcctcaataaaatcc | chr14:296600.f | gcttcaatgaacatcactacgg | chr14:296600.r | cagccataaaagttgtgcaaaca | 389 |
| **LIP** | chr14 | 296600 | no SNP | aactggaccaaggagatgtagaaaatttgaatagactgatcaccagtaaggagatcgaaacagcaattaaaaacctcccaaaaaataaaagtccaggaccagatggcttccctagtgaattctaccaaacattcaaagaagacttaatgcctatccttgtcaaactcttccaaaaaattgaagaggaggggaggcttcctaactccttctacaaagcaaacattatcctgataccaaaaccagacaaggacaacacaaaaaaagaaaattataggccaatatcactgatgcacatcgatg[c/t]aaaaatcctcaacaaaatactagcaaatcgaatacaacaatacattaaaaagatcatacatcatgatcaagtgggttttcattccggggatgcagggatggttcaacatgcgcaaatctatcaacgtgatacaccacattaacaaaatgaagaataaaaatcacatgatcatctcaatagatgcagaga | chr18:56233143.f | actggaccaaggagatgtaga | chr18:56233143.r | gttgaaccatccctgcatcc | 408 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | aagcatttgacaagatacagca tccatttttatgataaaaactctaa ataaaatgggtatagaaggaaa atacctcaacataataaaggcc atatatgacaaacccaccgcaa | | | | | |
| **LIP** | chr18 | 56233143 | SNP validated | ttaagacttttttttggtgaggaa gattggccctaacatccatgcc aatcttccaatttgtatgtggggc actgccacagcatgggttaata agccatgtgtgggtctgcacct gagatccaaatctgtgaaccct gggccgccaaagtgggctgca caaatttaacgactacaccacc aggccagcccaagactttctctt ttgaacccatttagcacatatttg agcgtctggtatattgccatgca ttcttccaggtactagagacaca gcaggaaacaaaatatcgccgtt ctcctg[g/c]agtaatgttctca ttgaggacaagacagacaaca atcaaataaaatgagctacattt atagggcgggggaagtggtt aaattcatgtgctctgcttcagtg gcccagggttcatgggtatgga ccacacactgctcagcaagcc aagctgtggcagcatcccacat aacaaaacagagaaagattact acaggtgtcagctcagggataa tcttccccaccaaaaagaaaaa aggtttatgtcagattgagggat aaagagaaaaaaaaaacatgg gtaaaggatagagcctggagg ga | chr2:13626700fwd | CCAAAGTGGGCTGCACAAAT | chr2:13627064rev | CCTGAGCTGACCACCTGTAGT | 365 |
| **LIP** | chr2 | 86268 | SNP also in parent | ggggagggctgggagagag taacacagcgggtgtttgagtct ggctctgggcaggggagggga cagtgacagccttggactgtgg ggacaggacggtctctctgaga tgggaccttagagttgagacctt gatgacaggatggaggacaga gcagctagtcctgcctcagggg gcgggggaggctccaaggag gaggggaccccggatctactc ctgaggatgcttggaatgcacc aggtggagagaggaagagaat ccagacaaggagacaggagc aaaggcccagagcggtgcag gtc[g/a]catcctgttggggaa tgtgtacgtggtcggaagtggg gctggaagtgccctggggctg catgtgacagacctgagatctgt gctaaagagatggggcaccac cctaagggaggggttcttccga aatgtttaagcagtgcggagcc gcccctctgaggagcatgcgct ctgacgcctgaggccctcca gcactgggtctctgtgagccgg acccgcagagactgttggcttg agcccagtgaaagctggacgc taatggacttgtactcaaataga ctgactgtcatgaaatcctggg g | chr2:26636712.f | ggaggacagagcagctagtc | chr2:26636712.r | acaagtccattagcgtccag | 428 |
| **LIP** | chr2 | 26636712 | SNP also in parent | gtcttggggacaggaaaccttg ctccagttcactgtaagtgtgct attcttggtgttgcctatgctctg agatttgaaattgaaaaactcag acttagaaaaacttaaaacaaat taacctacgtaaaactagctttttt ctcttggtgcacatttctgtgaat attaactcttgtgtgggtttacgt agctgccataatgatcaggatat agaacagctccatcacccccaa | chr22:23971258.f | ggaaaccttgctccagttca | chr22:23971258.r | ccctgcagtttgtgaggaac | 515

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | aacccctttcttcccgctcttttata gtcacaccctccccactcact gatctgctctctgtcactacc[a/ g]tttgtcaaagacccgggctgt ggaaacgcaaaagctttctctct gcttctgccaaagttgcccttcct tgaagccacagggctgcatgtt ccactgccccaagcagggcag gagtcgtggggcgtgaggaaa cgtggaaggacacgccgctca catttggagacagtgttttccac gttggtgacgtcaatgccttccc tatgtgttcctcacaaactgcag ggggtgcccctttgcttcccacat atcagtgtggctgtggttgacag cctttactggtatcatttttgtgaa tcaaggg | | | | | |
| **NOR** | chr1 1 | 7649446 | SNP also in parent | | gcctcccctggctcatccaaat gctccaagatattaactgagac aaaagattgagacagtagaaac agatcaggcaaatatgtccatg aaactctatcttaaggagaactt gacttctttctgtgtctttatgatgt aaatttgctgctcccatcttctcta ggacctgagagccattctttgaa atacaaatgtttcttagaaactag ttcattttgtattgcacctgattcat aacaaaaatttaacgtcttgtctg tgtctacctgtgtgtttatgtgtgt ctagatgtatgttata[c/t]atgt gtgatattttaccctggatggta tggccaaaattaaaagctctgttt aactgtcttaaagcaagtgcttat gtaaattctaaatgtagtagaaat taacccaatgtcttttaagttaac atgatatagattaatctttggtaa atgaaaagtttaagtttgttggttt gattgaaaaaaagacatgtcctt agagttgtcggtgtttgactatg atgcagacatgcagcctgggtt tactagtcagacaagctcatgtt gtctgtcagaaatttgtcagcaa aataatagcttta | chr11:38 548898.f | gctcccatct tctctaggac c | chr11:3 8548898 .r | actagtaa acccagg ctgca | 405 |
| **NOR** | chr1 7 | 20005831 | SNP validated | True deno vo variant | actacctgatccctccagaattt ggaaacctagtaagtgagtgag tatgattatttcattgcaatatagtt atttgcataagttcaatgagaatc tgttctccttgtaacaggacaca attggaaacattggttatcttaag aaagctttgactggaatgtcata tttgagagaaacatacagactc agatatgacaatactgcctttga ggaataaggttgactttctggaa ataaagccacttggaaatatgg gcctggtaccttatttacagaga gagattccagcaatcttacctg[ g/c]taagtaaggaagcttgctt atttggcaggtgcaaggagcct caaaatattttggggaatatcaa gagaagagagaaattcaccga aatctaaggtattgcgggcaag gcctgatggtacaaatccttggc ttggctttcctggcctcaagagg tctttaaagttcaatctgaggttc ccataaaaaattccagcaaag caaatttaagagcctgtatgatc aattgctattcttgttgcactcatg taaataattggcccaagtttattg aaacgaaacttattttgcaaacc aattag | chr17:45 65574.f | tggaaaccta gtaagtgagt gag | chr17:4 565574. r | acctcttg aggccag gaaag | 435 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **NOR** | chr18 | 21015205 | SNP validated | True_denovo variant | gctggcccttggtatacccttttt atcctcagtcttttgcaaagagat agagaaggtcagtattgtagca atgtggcatgagagagagcaa gctagggatccaactgcttcata aatacgatatggcctatttctaat ttttttaagcacagttgttcccccc attcatctcacttctagagattct gtcatgtaagtcagattaatcctt ggcttttgctatttccagactgca attaaatattttctttgatttgtttaa tccttaacctcttgtccattggtta atggcttccaaact[c/g]ttgac accattctgttctttcccattagct taatgtgttaaggagtcaaaaag attttattgaagttttaaagggaa gaaaagtatattcatgtttccact ctgccttctttacctggaagttca gctaatttatatttattatgtaaata ccttcctaagtgaagtatggaaa ttttttgccaaaattgtagaatgt tttctattgtgtgttagcatgctat gtgcacttaacaacatgtcgctg ggtgaaataaaaacaattatgc ataaatgcttcctttagtaccagc tggctat | chr18:21015205.f | gagagcaagctagggatcca | chr18:21015205.r | cacccagcgacatgttgtta | 471 |
| **NOR** | chr8 | 31590451 | SNP validated | True_denovo variant | gggccaaaccccccctccccc tccgttagaaatggtactacatta tcctactattttatgtaaaactgat cttaaaaataaccgctagatgg caataaggcataaaatgtaggt ctggcaagaggacacccattaa tagttcaaaggggcaagaaga gaagctggctcctcacagatgc tcaacgcctccctctccttggc atccccgaaggatcctagaagc tgtggctcaccatggtggtggt ggcttcatggggctggggtaca gggacatgaggagaaaggcg cagcagagaagc[c/g]tcaga gagtgaaatgaacgtgagcttt ggaatttgacagccctgagtttg tctttgctctgtcactcattggata tgggatctcggatgaagtgcttg actttgggggcctcggtttctcc acccttgaaacggtaacattaat acctcattcatagggcgtgatga ggatcaactgagatgatgtatg aacagcaccttgtatgggaaag gtgcttaagaaatggctgctagt acagggccaagatttgttatctg aaaaccttggggccagatttgtt tcagaattcagaattttcaa | chr8:31590451.f | accgctagatggcaataagg | chr8:31590451.r | atcttggcctgtactagca | 470 |
| **NOR** | chr1 | 115297935 | no SNP | | ttatatatgtaccaccttgggatg cagttaaaatatctaacgccatt ctattttgaagaacagtctttccc attcaagacattttagtatttaata agactattctttcttgactatcatt aagggcttgttgagtaaatttagt caggacttctatatgtgatatga catcctctagccccaaagaagc aacaaatatagccataggtggc cataccggtggaacactgaatg agcccatctggccttaaagaga agaagattggcaactggtgtta gctcagtgtgaatccttccttg[c/t]atccaagggaagctccggg tgcagtgttttagccatccaggc agtagccaaggctagagatttg ttccacatgaccaccaagttcca ttaggagccacccaataaatgc ttggcctttgagaccagtctgttc caaatcaatcactttctttaagtat gatagtattctgaaaaccttaaat | chr1:115297935.f | ctctagccccaaagaagcaac | chr1:115297935.r | gaccaggtctgtattctcgac | 396 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ggaacattataaaatag gtatac agtttaagcattacaaaggtttaa atgaggagctacaaggtcgag aatacagacctggtctggagtct tgggacagctgtctacaacaga tgccg | | | | | |
| **NOR** | chr10 | 21323697 | bad sanger seq | gtccacttctatccatcctcccaa gggaaatgagtgtacatgtcca tcaaaggccatgtacaaaaagg ttctaacagctttattcacaacag ccaaaagctgggcttaacacaa atgtccatcggcatcaaatgag ctcaagaaacagtggtatattca tacgatgaaataatacactgtag tcaacagtcagccacagacac gctcaagaatatggtgagatttt acagacaaaatgttgaggaaaa gaagacaggtacgaaagagtc ccatgctgcaggattccatatac tgaaattc[c/a]tagcaccaaa ataactaattagtgatgataaatg tgagaatattggtttctctaacca tagaacatatggactgagacgt agtatagctcgtgtcacatggaa aaactgtctagaatatggaaaa catttaacgtatacaccttgaaat gggtggtagtttcacaaaaatat atatctgtatttctatatctgcacc tgtctgtctgtctgtctatctatat atatatttattgcaccacaattaa gactagtggactttggaaccact tcaataaaaaggcaaaaatagt gcatttgca | chr10:21323697.f | tccatcggcatcaaatgagc | chr10:21323697.r | agacagacagacagacaggtg | 400 |
| **NOR** | chr11 | 38548898 | bad sanger seq | tcccttgcccttgcttttcatttcc ctgtttccttttctccttttagacaaa cagggtctgcagggagatgga ttctccctgggttttgggtttgga gctctgctaaggtcttgggaaa gggccacacttgccggcaggt ctgctgtctcgagcaggtgctg aggcttctgtcgctccagaagg acacttttttggaccaaggatgtg ggcctcctggccaggagtctag gtggatgaaagcatcagcagg acctccccaggtttggggagag gagacctgttcatgcacaggtg ggccgggg[c/t]gggggaag ggagagcaggcggctttgata aggatgaaacaaaatgaatgtg ttttcaaacttttccccaggcccc agcaggattctgcaaaactgta atcccattttgtaaatagacgga attgtctctctctctttttttttaac ccactggagatgcaaggccag ctactgtctctcctctctgtctctc tcaatacaaatgacaaaatggg aagggaagctgcccgtttgtcc cgatgagcgatgcggtaagtg aggtgcagggcggtgtcggtc ccagaggcctgttctcttggt | chr11:7649446.f | ggtctgcaggagatggatt | chr11:7649446.r | agtagctggccttgcatctc | 423 |
| **NOR** | chr17 | 4565574 | bad sanger seq | gtaaacataacctgtgcacatttt gtcaaatcagtgacatttctagg taggaagtttgaaaaccatcaa agattctaggacacacaggatg gtggtttaggcaaatctttgctg gatgatcatttatcaattcctttcc cctttgtcattatgaaattgcaag actttctgaagaatcattagatag aggaacatactaattccattcctt atagaatgacaggtaggataat gcagaatataagaaaagtactt attctccagaattggcaaaatgg | chr17:20005831.f | caggatggtggtttaggca | chr17:20005831.r | aagccaacctcgtgtgtttg | 416 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | cagcaaagtgataaccatctac[c/t]ctgactggctccctgtcaaaagttacatagccttcctggcttcaccctcttacatatctgagtgatcttttttaaatacaaacctggtcaaatcccttggcagaacacccccctcagctttgcatccttctggaattgtaacctgtgtgctcagccatttggaactgccattgttcccaaacacacgaggttggcttacatctccacgacttaatgctgtcccaatgcttgtctgtctgcttagggaacacctgccaatcttaagagaccaagctcaaggtgatgctgtgaagccttatc | | | | | |
| **NOR** | chr29 | 1231244 | SNP also in parent | tgttccaggaaagggacacggggagagacagtaggcgtccacaccaggcgagacaaaaccaggtgagcccagaggagggaggtagagaaccaggcggactcccaggagccaagggcggtctcccccctccctctgacccaggacctggccggccgtccctccccctcctccttccgaccctggacccggccggccatccctcacgctccctcccacgcgtgacccaaccggccgtccctccccgccctccgacccgggacgcggcctgctgtccctcctcctccctctggttgttcctctcc[c/a]gcacctctgagagcaggggagcccagacacacgcagtgtgcacagcaggcagcctctctgagcctgagacagccgccagggctgccccccgccgctgggtagcggagacagaggggatggagggaccgagaagcagtgtgcagcccgcgagccccgcgccccggccctcgaggacagcgccccgcaggcctcagcggacacggcgggccagggccgcgccgccgcccggaaggctgtccgggaccccatgcccgggctcagggcgcgtctccctgggagcggccctgcccgcagtcctccg | chr29:1231244.f | gggagagacagtaggcgtc | chr29:1231244.r | ccctccatcccctctgtct | 426 |
| **NOR** | chr7 | 64671737 | SNP also in parent | tcacagtcccaataggaaaaaattgccatgttaaaattagacaattcaaggtgggttcaataaggagttgatctaaaatggtgtggacacgagtgcacagtaactgagggctaatgacagtgagatgcgccatccttcagcctggagtgaggaggagggaaatgttactggaagcaggaaggaaagagaactatagagccaactgtcttagaaggttagtgatgttggtagagggacatagcctgtctgaggccaccctacagggagggagagtgccacaacttgactcccttccttcatt[t/c]ccctctcatctcctgctggggcacccctggccaaaccaaccagaaactagaggaagggagactcgatgttaccgacaaagggcagagagcagagtggaaagggataaagtgaatccgcaagggcaagcaaaagatgttcagcccaaacaaacccttttatttaaagcccatctgtttgctggaaacacataggaatgaaattgtttccaccctggagggagctacagtcaacagagagtttggagaaaagtaagctacaaaggagttccaaccctgagaatgccataaggaatttattt | chr7:64671737.f | ggctaatgacagtgagatgcg | chr7:64671737.r | tccttatggcattctcagggt | 487 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **NOR** | chr14 | 473741 | no SNP | ttttaggctggttaattttaaaactacagatgagaagcaggctccaaggggtggaatttgcttgctctttgttgggaggcatttgcatttgtgggggaacctccatctgtggagatgtctccctctctgtgccaggggggaagagggatggccttgtctctggaaacttttaatggggaagcaagaacttaagttggttactttctggcaacctcatgtaactgatttagggtggtggcttctggccttttacttaatctgatttgattcttatctaaaagatgtgggaccacccaatgg[c/t]cagaccccacctgcactgataccattttaacttttttccatgttctttcctttgtcttgtaaagagatggctcacatacctatgccttaaatttagccttactctccacccacgttggcagcagaagcgggggcagcacctcctcctgcccatgggtcctgtccccatgctcttccacactattctctaaaataaaagagcactaccgccagatcttgagagtacaagaaatctttctttcgactcctcggctcactgaccccgcatcacactgacaaataaaaatggcaagtaataggat | chr8:179 9905.f | gggtggaatttgcttgctct | chr8:17 99905.r | tcaagatctggcggtagtgc | 473 |
| **NOR** | chr27 | 596466 | no SNP | aagcagaaagagaaattaagaatacaatcccatttgctattgaaacaaaaagaatagaatacctaggaataaacttaaccaaagaggtgaaagatctctacactgaaaactgtaaaatattgttgaaagaaattgaagaagacacacagaaatggaaagatattctgtgctcttggattggaagaattaacatcattaaaatgtccatacttcctaaagcaatctacagattcaacacaatccctatcaaagttccaacaacatttttcacagaaatagaacaaagaatcctaaaattta[t/c]atggaacaacaaaagaccccgaatagccaaaggattcctgagaaaaaagaacaaagctggagatagcacactccttgattcaaaataaactacaaagccatagtaaccaaaacagcatagtactggcacaaaaacagacacacaaatcaatggaacagaatcgagagcccagaaataaacccaaacatttatggacagctaatattcgacaagggagccaagagcatacgatggagaaaggagaggctcttcaataaatggtgttgggaaaactagacagacacatgcaaaagaatgaa | chr6:327 0286.f | ctgtgctcttggattggaaga | chr6:32 70286.r | ctgtctagtttttcccaacacca | 420 |
| **NOR** | chr10 | 13100245 | SNP also in parent | gccacacctcttatctccacaccatatctcgctctctgttttggccatttttctgtttcccccggctcataaagctgcacaaaggcaggtgcagtgggatctagggtgcctggcacacagcagatgctccataaacctttgaatgaagcctatgggtccccctgtgccagctctactctcagccctttgcatggactatcgagtctatccacacgaccacccagcaaggtaggtagtattattgctcccagtttcagaaggggaaaacaaggcacagaaagaagaagagttgggttaga[g/t]ctccaggtgtgaaggaactgggcaggaaagcaggacaatccagggtggggaagcctgggccccagcgcagggccatacctcttgcacagcccacgcacctacgtctaccaacagaa | chr10:14 151126.f | caggtgcagtgggatctagg | chr10:1 4151126 .r | atccaaccaagctgtcctga | 471 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | gcgtctcctgacacatctctcca gcctcacacgtctgggacccaa tcgcggtccattcaacgggtgt ctgagccgagtttcgctgcaact tcaacatcagtccccgttgaca ggtcaggacagcttggttggat gaagcagggtctgggaaagcc agggaaacagccgtgtctgtcc ctcc | | | | |
| **NOR** | chr10 | 13100245 | no SNP | tccatatttatcagctctcctgttc agggagtctatattcttcttaatct cctccattgtgtttttcatctccaa catttctgattgcttcttctttatag tatcaagctcttttgtgatgtagct cctgaactcattgagttctctatc tgtattctcttttgactcattcagg ttttatttttttgtggaagattagcc ctgagctaacatctgccaccga ttctcctcttcttgctgaggaaga ctgaccctgagctaacatccat gcccatcttcctctactttatatgt gggagg[c/t]ccaccacagca tggcttgacaagcgatgtgtag gtcctcacccgggatccaaacc agcaaaccccaggctggtgaa gcagaatgtgcgaacttaacca ctgcacctccgagctggcccct cattgtgctttttaatgatagctat tttggattctttcacatttaggtta cagatttctgtgtctttgggattgt tttcctggtacttgtcattttccttc tgttctggagacttaatatatttttt catactgctatatggtgttgattt gtgcctccacatatagagtttcct tcctg | chr12:17455667.f | tgatgtagcctcctgaactcattg | chr12:17455667.r | ggcacaaatcaacaccatatagc | 463 |
| **NOR** | chr10 | 14126430 | SNP also in parent | gagccctaaccggcaccacat ggacagcaagcacgcattagt cactgagccaggtaagacccat atgccgcctcctggagtttgctg tccaggtttggagaggtcatcat gcaaggaggcaggtgccagtc accaagaagggtgtgaccaag tgccgagaactcaaagctgttg gagaccccaaggaaagattttt agagaagacggcattttgaaca ggaccccagtagaccaggcaa atttccctggcagagatgggag gagccaccacgcagaggaact ggtgtgctcacagacgcc[c/g]agaccaggaacctttgtccac aagaatgaattctgtcctcctgg aggtgggaggtgaagagggtg gggaaggtaggggagtggctg gaatcacattgctgagagctgt gagtgccaggctggagtctga gccttgccccatggacagcag gaagccgtggaaagttttggag caagacaagatgaaagccgttc gtgagcactctgatgtctgttctc agacccaggtccccaggcccc tgcaaattacccttccagcagtt cacagaagagctgtgtttcagc cccttgtgtatctctg | chr12:17467671.f | gcaagcacgcattagtcact | chr12:17467671.r | gctcttctgtgaactgctgg | 550 |
| **NOR** | chr12 | 17455667 | no SNP | ctcaccaagaaaaaaagagag aaggctcaaataaatataatcag aaacaaaagaggagagattac aatggacacctcagaaatacaa aagataagaagagaatattatg aaaagctatacgccaacaaatt ggataatctagaagaaatggat aaattcttagaaacatacaacctt ccaaaactggacaagaagatgt | chr10:23731364.f | aggagagattacaatggacacct | chr10:23731364.r | atgttgaaccatccctgcac | 538 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | agaaaatttgaatagaccgatc accagtgaggagatcgaaaca gaaattaaaaacctcccaaaaa ataaaagtccaggaccagatg gcttccctggtgaat[t/c]ctac caaacattcaaagaagacttaat acctatccttctcaaactcttcca aaaaattgaagaggaggagag gcttcctaactccttctacgaag caaacattatcctgataccaaaa ccagacaaggacaacacaaaa aaagaaaattacaggccaatat cactgatgaacattgatgcaaa aatcctcaacaaaatactagca aatcgaatacaacaatacattaa aaagatcatacatcatgatcaag tgggtttcattctgggggtgcag ggatggttcaacatccacaaag ctat | | | | | |
| **NOR** | chr28 | 20549421 | no SNP | ccttctgccgtaaatacaacgac aatgacctgttctatttagaggct ccaccttcgtcctgaaacagaat atccaggtgaggacaaccagg tgtggaaccaaaaaagaaaca cagataaacacaaagagcatct gaccaaggcatcttttaatcaaa ttgctgaaaactaaagagaaaa tgttaaaagaagccatggaggg ggcacattacatacaaaggaac taagataagtatgtcttcaaactt tttatcagaaaatatggaagcca gaagaaattaaattacatctttaa atgtac[t/c]gagagaaaaaaa tattcaagttaaaattctaaatgc attacaaatattatttaaaaacaa gatgaaatgaagattatttaaaa aaaaaagaaatcttgtcaccag aagacatgcttgacagtagatg ctaaaggaagttgttcagggtg aagaaaaattacactagatgga aaactgaatctccacaaaagaa taaataacataacaagttatatgt aagtaaatataaaaggatatttct cattgttcaagttcttaaaaagaa aaatatttgtttaaataaaagaaa agaaaaacaaa | chr20:32276595.f | acaacgacaatgacctgttct | chr20:32276595.r | ccctgaacaacttccttagca | 435 |
| **NOR** | chr20 | 32276595 | no SNP | atattgcatgttctaaaaatggct gcttcgccccttctagaggagg agaatttatgatgttaatgaggc aggttactttagtacaactctgcc tgtccgctcaggtgtcattcttat ggttgtggtctgggctggtttgg gccagttcagagcagcagggt ccagggcagttgttgcttagagt ttcttcctggaacacagctgaaa ccacaaaagctcaagactccaa agtcattagagaaagcaatattt gcacctttcaacaagacaagag aattaggtcaaacacaaaaggt taga[g/a]gttaggcaagaca agaagagtccatgactagtgag aagatgatccagaatgaaattta gcatctggtgcaagatctctgg ggaggcccatgaaactctcacc aacaccatgcttccttcctcttac ctctgtcccatttggccagcaag agctattctcacagttggatcag atgcatactcttcaattcagcca cttggtggcaataccacagcca ttgatgacacaccacagaggtg ccatatctccaggaaacaccatt | chr23:40878726.f | agtacaactctgcctgtccg | chr23:40878726.r | tgtggtgtgtcatcaatggc | 457 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | cctccacttgaaacttggtgctta ccattccctcctcc | | | | | |
| **NOR** | chr22 | 36838245 | bad sanger seq | taccaagaaggcggagatcat aaccttgcttacaccttcggagg caccggaatggtaggagagaa catcgctccctcccctagagtct gcgatcgctgcggtgcaggtc gggaaggagtgggggagggg ccgcgcgaccgggggatcatc caggactcctgccgctgattca gtggagacccgctgatgggggg gaaagcttccgtccacgggga ccctataaatcaagggccttgg gagaccagagaacagaactga tctgaacccagactggtgcgtg tgagtaacagcccctcccccc[t/c]aacaaagccagtgggtgc aggcatcttgccccgaaggcg gagagctaacatgccgctctca acccccatctagtggcgacag gctgtaactgcaactgaattcta ccaccatgagaaaaaaccgct cctctaccatccagcaatttataa aagccccagaccagaaggaa aacaataaaaacacagaattaa gtcctgaggacttggaattaggt aaactaagtgatgatgaattcag agcagctataatcaaaaaactc aatgaggtagagagaaagata gagaaacaagccgagttct | chr8:49612506.f | tctagctcac tgatccgttct | chr8:49612506.r | cacactgc tggtcctc ataag | 408 |
| **NOR** | chr11 | 41656509 | no SNP | gtcccagagttccctaacactgt tctcattctgtctaatttttttctc ttttctgttctgctttggtgatttcc tctaatcttttctctagctcactga tccgttctttttgcttcctctactctg ttattgagtccctctagtgaatttc tcatttcaagtattatattcttcattt ctgattggttctttttttatatcttcc agttctttgctgatgtgctcactg tgttcatccattcttctccacatat ctgtgagcatcctcattatattttg tttgaattctttgtcaaggag[g/t]ttgctagtttctgtttcacttagtt ctttttctggtgttttgtagtgttcc cttgcttggaaagtattcctttgc ctcctcattatgcctctttctctgt gctgttttctttgcactaggtgag tcggctatgtctcctgatcttgga gaagtggccttatgtatgagatg ccttatgaggaccagcagtgtg cttccctctcgtcaccagaccat aagaaccaggagtgacccctttt gtgggttacttgtgttcttctgtg gcagggttgctcccactgcagg tacccagggagtgtga | chr8:49612506.f | tctagctcac tgatccgttct | chr8:49612506.r | cacactgc tggtcctc ataag | 408 |
| **NOR** | chr1 | 44584336 | bad sanger seq | atcagtgatactggcctgtagtt ctccttttttgtgctgtccttgcca ggctttggtatcagagtgatgtt ggcctggtagaatgtgttagaa agtcttacatcctccctaatttttg gttagctgaaaaggataggtatt aaatcctccctgaatgtttggta gaattccccaggaaagccatct ggtcttagggttttattctttggga tgcttttgattgctgtttcaatctct ttccttgtgattggtctgttcagat tgtctgtttcttcttgactcagcttt aggaggttgtaa[g/c]agtcta agaatttatccatttcctctaggtt atccattttgttggcatataggttt tcatactattctcttataatccgtt gtatttctgtggagtctgttgttat | chr1:155138900.f | tcagagtgat gttggcctgg | chr1:155138900.r | gacaaac ccctagcc agact | 429 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ttctcctctttcatttctgattttgttt atttgagctttctctcttttttcttt gtaagtctggctaggggtttgtc aattttatttatcttctcaaagaac cagctctttgtttcattgatcctttc tactgcctcttttgtttcaatagca tttatttctgctctgatttttattattt ctctccctctac | | | | | |
| **HAF** | chr10 | 59172497 | SNP Validated | True denovo variant | ccaagctttccccactccaccct gtttcttcacagccctaaaagttt gtccaaaaaagcaaagaggatt atgtcatactcttcatgagaagtt ctgttggctcagcttccttctgga taaaatctaacattagttagaac agacagtaagttcaacctttcttt agtggatctgtttgggttttttcc tccttaacatctgaccacttttcct atttggggtactccccacaggg ccctaccaccctctaaagtggct cagggaccagagtctccctctc cccaccctgtgagagccag[g/a]gtgtggggatatgaactggg ctcgactagtcgtgtgctcctgc ctggggccttggatctccaggg acacaaggatacaggagttga gaagttattctccgtgaaggcg gcagtgccaacctagggccag caggcacctgtggcagcatcct ggctgcacatccctggggcatg agtctggccatgttttcctgctgc caggcctcccttggttccagcct gctttccaaagctggttcttcag ctttcctgccgattgggaggtca ccagtatcccttcaatacagacc attcctcccta | chr10:59172497.f | actccaccct gtttcttcaca | chr10:59172497.r | gaaagca ggctgga accaag | 516 |
| **HAF** | chr17 | 1685198 | no SNP | | agttttcccatcaccatttgttga aaagactttcttttctccagtgta ggccctgagctcctttgttgaag attagctgtccatagatgggtgg ttttatttctgggctttcaattctgtt ccattgatctgtgcacctgtttttg taccagtaccatgctgttttgatt actgtaactttgcagtatgctttg aagtcagggactgagatgcctc cagctttgttcttctttctcaggat tgctttagcaattcggggtctttt gttgccccatatgaattttaggat tctttgttcta[t/c]ttctgtaaag aatgtcattgggattctgactgg gatggcattgaatctgtagattg ctttaggtagaatggacattttac ctatgtttattcttctgatccatgta catgaaatgtctttccccctcttta tgtcgtcatccatttctgtcagaa aatccttgtaattttcattgtataa gtctttcacttccttagttaaattc accccgaggtattttattcttttg ttgcgattgtgaagggtattgtgt tccagttctttttctgttagttcctt attagagtatagaaatgctg | chr17:16185198.f | ccattgatct gtgcacctgt | chr17:1685198.r | acaatacc cttcacaat cgca | 434 |
| **HAF** | chr2 | 100761774 | SNP Validated | True_denovo variant | aggcacagataattttaaggaa atcaattccatttattcaacttttac ctactcttgcctaaaattgagctg cctgaaatatgcctgaggttga gatgtcacaacaattccacttct ccaactccctttaaaaatcattct tttcccaagtataaaaagacaag cttcttctcacctatcccagatctt actatggcaaattgctcctgtac atgaaaagcctctaatgtaccat aaaaaggaacaattctaaatatt gatggcaatatttaatttaaccaa | chr2:100761774.f | tgcctgaggt tgagatgtca | chr2:100761774.r | ccagatga agtagggt gcaac | 466 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ggcatgataagcctacag[g/t]aagacttcctggctttctttatcttacaaatatttctgttaagagaaaaccatggtggtagaaaatgggtatttgggtctgtgttggggtgctctgagttcagatatattatagagtggggtggcaggagttatgatcatttttgtctaatgtcatttcctcttccattttctgatgatgtcaaagaatactattgagatagcaaaaagaaatattgaaggggttgcaccctacttcatctgggtaacaagctcacagcaaggtttcatgctccatatatcaatctcagaattcaaata | | | | | |
| **HAF** | chr3 | 53100714 | SNP Validated | True_denovo variant | tgtaatccaataccacattattccagcattggcccttgggagctcttccagttgcctcccatgtccatccaatatacacccatcattgtggggttttttatgagcactttctaacttttgtcaggctctaggctcatcttggacattccctgccccagccctagaattagctatttctccaaggagccctgattccttttattgaggattcatattagaagcaaacatctggattcagggagtacttgttgctaccagtgtgtcactgctctcaacagccttttctctaactcacaatt[c/a]tacacattaatcatgagctgggctcagctgcatggttcttgtggtctggactgaattcagccgatttcatctggacttgctcatacatctatggtcgtctggtgggttgtcacaagtctttaggataaccttggttaggacacttgagagggctgtgacctcttctctacatggtcttctatcctccagcacactaaactgagcttgttcatatagtctcaggattcatagcagcaagaacaaaaatgcacaaagtatcttgagtccaacgatggaaacttacccaataatgcttgagcta | chr3:53100714.f | ctcttccagttgcctcccat | chr3:53100714.r | tcacagccctctcaagtgtc | 419 |
| **HAF** | chr4 | 44983704 | no SNP | | agacttgtgtttcacaaatagccataaaaaggatcagccccacctcccaaagcctgaaacaactgagtgcctccatgccaagggctagccctcctcagctgcactcctaagagaactgacatcagccttgttggcctcaggcctatcacaactgtacgcccctgagcctagcaaccagctacactgggtaccaacccaattaagaggacaattgcaataagagtgtgctaacaggcgggaccaagatggcagagtgagtggtcttctttgtctctccccccgtcgaatctacaactaattg[g/a]acattcaccaattaacaaaggatatccagatagcatctcaagacgcctaagagtctcatactgctatacatcggaaggcagacggtcttcccccccgggaggaggtggaaatagggtgaaaactctctgaccccaaccccgaccctcagacagcctagttcctgcaggaggctctcttccagcagactcccccacagcattgccacacacccaagggtgggagtgtgcactcaccagcagagcgacggtggaaacaggtgacaacagccctactcaagccccccgtgattacacataagc | chr4:44983704.f | cctcagctgcactcctaaga | chr4:44983704.r | tgagtaggggctgttgtcacc | 487 |
| **HAF** | chr5 | 21907765 | bad sanger seq | | ctgtttctgttgtcactttgaatagatacctgaacatggtttcgcagtaaactttgtttcccatctgaacaaatgttttaaagcttttaatatttttg | chr5:21907765.f | cctgaacatggtttcgcagt | chr5:21907765.r | tccagagtgtcagagtcctt | 556 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | acaagcttccccccacacacaa atattcaagtcctgaataaaagc tttcttttgacttggaagaggag ggaaggtttctctggggattttc agagggcccctgggacttctca gagaaatttgctctctttctataa gaaaaatgctaaactaataagg cttattttgtatgttaaattacatga gaagcattgtcaaataa[a/g]t aatgataaactttcttaaatggta ttatgtaaatgagtgttattgatat aaggattttaaaaattatctgatc tgattgtcagccataattgtggtt atctcaaaataattctggttagta tctcaaaatgttatatatcacaga catggtcaaatgtctttgtcagtt accatttttgaatgttttgttatttg cagagagttgctgttttactttaat gtttttgcaaatatgtttcatctcct gagaaattcatggaaaggactc tgacactctggaatataggtttc caataaac | | | | | |
| **HAF** | chr5 | 53320088 | SNP Validated | True_denovo variant | ggtggccacaggaacatttttta aagcaatttttctttccttttcatg tataaagagcctggtacagagc ctaacatataatgaacattaatta aatggtagctattgttactattcta gaactcacagaaagccatctaa atctagaatttatttgaatttcacc agccttcgaaattctagttatttct gccacaaaaccttcccacacct aattggtaaatgattagctaagg gatagtccagatcactgctatca ttcattttcaatctgacagaagca caattttggagcgagt[t/c]tttt ctctgataaaatatgtttagtgga gtaggagtcctgtggaagtatc ctacagaactgtgcagtggaca tccgtcacctgtaactgccccca tcgtttcaactgcacctctccac attttcatatttcccacattgtgaa tcctaccttcccaaataatttttttt catccttatatttcctagcctcctt gcttctagaaaacagacttctga cttagatcttataggtcagttacc tgagcacagtttggtttggaatg agtttcatggggttgggcctgg gtactggcgtgggc | chr5:53320088.f | caccagccttcgaaattctagt | chr5:53320088.r | ctcattccaaaccaaactgtgc | 406 |
| **HAF** | chr7 | 91483698 | SNP Validated | True_denovo variant | ctccaaatctccattttcaattcta aagggagggaacataatgggc gttttttgtgctccatcatccttgc tgtcactccggtctcagaggctt cagccagtgatattcgttagaat cgtttttgactgagaatagtgag atctcaaaaaaataataacctaa acaaagtacaagtttatttccttt gcattagagaagtttggaggta aaatggcttggaagatgtgtca aggaagcagtatccttcttaatg tgacttctgcgtccagggtaattt cttggcttggcacttcagtcat[t/g]atatttactttccaggcacca ggagggagaaagaaaagggga aacctttgtgactaggttttccgt gcacctggtgttcattcaccttat tggccggaacttagtcacaatg ctaagctgcaaaagaaatgtaa tcatttagctgagcaccatgctc ccagttacaaattgggatttgttt actaagaaaggacaaagagaa tagatactgggaggcagagag caatctctggcccaatttcataat agagctgccatcgctgggttca | chr7:91483698.f | tcatccttgctgtcactccg | chr7:91483698.r | cagcaactggaccagtgg | 537 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | acgccccccactggtccaagt tgctgcggtt | | | | | |
| **HAF** | chr1 | 120334435 | no SNP also in parent | tgaggaaatacattgtttgtaga agttaaaaacaattaaactgaca aagttggtctgctctttgttatccc cacgtgccagcaattctcaaca atatcagtgacaaaataccaatc agagcagactactcccactcctt tctaagttgacctatcaatgtttc atgaaatcaattatttaaaatatg gaagaagaaaagataaaatgg agtgaacagaaaggagtgcat ctcacacagtatgagtagatact gctttgttaaacttttgtttcagag acagacatgcagatacacatgc a[t/c]attggattgtagcataaa atatacttcttactatttgtagcag ttaaaaatgctagcaattaaaaa tagccactgggaaagaataaa gaccctaatgaagtggtatgca ctgattaaagagtactagactag gaatcagaagttttgggcttttaa ttttggcttcgtcttcaaatgcgtt ttgtatcttacagatatgtatgtga aaatattttctttgttaatacaaaa gatatttaaagaattttccaggta tatatggaaaagatcttttcctta ggaaaaaaatattcttcacca | chr1:120 334435.f | ctttgttatcc ccacgtgcc | chr1:12 0334435 .r | acgcattt gaagacg aagcc | 424 |
| **HAF** | chr1 2 | 13784646 | SNP also in parent | agagtgatggaactgttctgtat cttgcttgtgtttgtggttatacaa ctctatgggtttgtcaaatctcgc agacttcacaccaaaaacagta tattttactttgcatgcaaatttaa aattcactatgtgtaatttaaaata atttttttaaatatttttgggtaaga aagacattgctatgatgctatatt cttaatttacaaattttacttaaac caatttatgctgttaatcataaatt agcaaaactttaatagagcctcc tgtaaagagcttctagaatttgtc caaattatta[t/a]ataattggct ttcctgttactgactcccagtgg ctcctttgttctagatagtttagcc acctcttcatgaacctagcatct cgggatcacccaagatgatcat aagttgccaaggactacaggta caagaaactgccctcctcagct ccatgctacacatcattagtatg gtcattgcaaggtgcatcctgg cctgtcctgaaaatagaggctct cagagacctctatcctttgctgc ccatttgccttgtgtcatgaaag cagcattcccacaaacttagcgt aggcacagaggcctcc | chr12:13 784646.f | gtcaaatctc gcagacttca ca | chr12:1 3784646 .r | gcaaatg ggcagca aaggat | 489 |
| **HAF** | chr1 4 | 4827882 | no SNP | gggattctgattgagatagcatt gaatctgtagattgctttaggtag tgtggacattttaactatgtttatt cttccagtccatgtgcatggaat gtctttccatctctttatgtcatcat caatttctttcaagaaagtcttgt agttttcgttgtatagatctttcac tttcttggttaaatttatcccaagg tattttattcttttttgttgcgattgtg aatggattgagttcttgagttctc tttctgttagttcattgttagcgtat agaaatgctactgatttatgtatg ttga[t/c]tttataccctgcaact ttgctgtagctgttgattgtttcta atagttttcctatggattctttggg gtttttgatatataagatcatgtcg tctgcaaacagtgagagttttac ttcttcgttgcctgtttggattcctt | chr14:48 27882.f | ttcttccagtc catgtgcat | chr14:4 827882. r | caaaccc acagcca acatca | 525 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ttatttctttttcccgccgaattgct ctggccaacacctccagtactat gttgaataagagtggtgaaagt gggcaccttgtcttgttcctgtt ctcagagggatggctttcagtttt tgtccattgagtatgatgttggct gtgggtttgtcatatat | | | | |
| **HAF** | chr16 | 2815931 | bad sanger seq | tgctgtcgattgctttaggtagta tggacattttaactatgtgaattgt tccaatccgtgagcatggaatat ctttccatttctttgtgtcttcttcta tttctctcaaccatgtcttatagttt tcagcatacttggtctttcacttc cttggttaaatttattcctaggtac tttattatttttgttgcaatcgtaag tgggattgtattcttgagttctcttt cggctagtttgttgttagtgtata gaaacacaactgattttcgtgaa ttgattttgtaccctgcaactttgc tg[t/c]agttgttgattatttctag tagttttctggtggatttgttagg gttttctatgtataaagctgtgtca tctgcaaatagtgacggttttact tctttggttccaatgtggatccctt tgatttctttttcttgcctaattgct ctggctaggacttccagtactgt gtcaaatgggagtgacacgaat gagcacccgtgtcttgttgctgtt cttaggggatggctttcgggtttt caccattgagtatgatgttggct atgggttgtcatatatgacctttat tatattgaggtatttt | chr16:2815931.f | tgctgtcgatt gctttaggt | chr16:2815931.r | gtgctcatt cgtgtcac tcc | 497 |
| **HAF** | chr20 | 29311733 | bad sanger seq | taagcaataaaaagattattgat gagatattttacattctttttctcat tccttacttttcttgaaattattgtt cttagccgccaaagggtcaga aattccttaattttatgtcctgcca tgtggtgactttcatccagatcc agctatcctaggatttgctgacc acatcaaaatcacttacgcctca aactcgatgcaggagagagtg agcacagagacagagctggc gtaggaccccacctcagcttcc tccctcctctcccagctttactcc acccaggagcccttcatcctcct [c/t]caatctccaggattccatg aatctcacgaatggggagcca gtcatgagatatataaacccag atccccacaattcccctgcagg aagtccttcctgaagtctgccca ctcccattcctgctggagcctca gtgcttccccaatagagcccca ggtctgtgcagagggtgatcta gagcactcttcctccctataacc acagggctcactccctcacctc cttcaggacttagctcaaatgtc gcctctcaggaaagcctcccct gaggataatatttaaaattacaa ctagttccctcat | chr20:29311733.f | tcctaggattt gctgaccac a | chr20:29311733.r | aggcgac atttgagct aagtc | 400 |
| **HAF** | chr18 | 52005 | no SNP also in parent | tcacctgggaagtgggccggc tgtcaggccgtctgctctctcca ggctcactgttaggggctcctg ctgctggtggtccctcctcgcac cagatgaagatcatgcctcagc ccatgactccacagctccaacc ccttctctttcagtccccagggc aagccctaagtggtcagactgg tctggggaccctgcacagaca gcctgaggcaaggcctcccag gggccgatgtggggctgaccc tgtccccaggccccttggcattg ggcatcagctctgagtcatcca | chr18:52005.f | agccctaagt ggtcagact g | chr18:52005.r | atggctgc agagttgt cct | 406 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ctcagggacagtct[c/t]ccct ctctgtgcagccagtcgcagtt ggggcaggcagaaagcggcc cagatggtcagccaggcaatg gtctacccaggctggaggcag ctagaatgtacccaggccagg cgtgaagggcctcagtgccac ctctgatgccatcaaggcccca ctgtcccctggccgggtggcct cctgggcatgcaggggcctctc cccagtcgcacagcttcctttgg ctcttgcaggtggttgggcggc aggacaactctgcagccatgcc ccacctgccgaacctgcgctga gctcccaggactcat | | | | | |
| **HAF** | chr24 | 540464 | no SNP also in parent | tccttttttctgtcttatgtcaattta actcttagaacagccagaagaa cctagaagaatagagggaaaa ttcatccccccctctatcctatgta gggattcaacttctaatagcaca ttctttccagattttgagcctctgg ccgcagtttatccatgtctgcata tgcctgaggagaatagtatttgt agtcggatactcatgaacaggc gttacagattacggacgcttccc cgtctgcggcctctgggagca gtcaagcccacccaggcttaat acagcagagaagggtgcaaac at[a/g]tgctgtatcgacagtca gtatctggagacagtaatgtga aatggaaaaaatgttgtgagcta ataggaaatatagataggtctct ggccacagttcctgacacagg actcctggaacactggtaaattc ctaagtgccaagagcaatagaa gcttcttttgttctaactaggtgac tctgggtgggctcctgtgtggct tttggatgcaggctggacacca gaaagaccaagctgtgattaga agcttgggatttccatgcctcctt tctcctccaacttctctagagag agagaggagg | chr24:540464.f | ctggccgcagtttatccatg | chr24:540464.r | agttggaggagaaaggaggc | 444 |
| **HAF** | chr10 | 7931916 | no SNP | cttgtcaaatgctttctctgcatct attgagatgatcatgtggttttgt tcctcagtttgttgatgtggtgtat cacgttgattgatttgtggatgtt gaaccctcactgtgtctctggta tgaatcccacttgatcatgatgta tgatccttttatgtctcgctgaat tcgggttgccaaaattttgttgag aatttttgcatctatgttcatcagt gatattggcctgtagttctccttt ttgtgctgtccttgtcaggctttg gtatcagagtgatgttggcctca tagaatgt[g/a]ttaggaagtgt tccatcttccctaatttttggaat agcttgagaaggataggtatta aatcctctctgaaagtttggtag aatttcccaggaaagccatctg gtcctgggggttttattctttgggat gcttttgattcctgtttcaatctctt tccttgtgcttggtctattcagatt atctgtttcttcttgactcagcttt gggaggttgtaagagtctaaga atttatccatttcctctaggttatc cattttgttggtatatagtttttcat agtgttctcttacactctgttgt | chr10:7931916.f | tgttgaacccctcactgtgtct | chr10:7931916.r | ctcccaaagctgagtcaagaag | 417 |
| **HAF** | chr22 | 24976376 | bad sanger seq | ctctacagaaaccttacaagcta ggagagattggaatgacatatt caaaactttaaaggataaaaatc ttcagccaagaatactctatcca gcaaaaatatccttcagatatga | chr22:24976376.f | agctaggagagattggaatgaca | chr22:24976376.r | tctggtggcctcttacttcc | 451 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | gggaggacttaaatcctttccag acaaacaaaagctaagggattt tgtagccacaagacctccacta caagaaatcctcgggaaggcc ctcctacctgaaaaaagaaaaa aagggagaaaggggtcacaa aatacagagtagggagactaat agatagaaccagaataggata gcaaatattcaac[c/t]atagca ttaggataaagggaaggaaatc accaaagcaaagacaatcttatt gctctaaccacaaactcacaac acaagttggaataagagatgaa aataataatttaggaggggaag aggaaagggattgaatcagttt aggctaaggaagtaagaggcc accagaaaatggactatgttata catgaggttctggatacaaactg cagggtagccactaaactaaaa aacagaacagagacacaaaac ataaataaggaaaaagctaaaa aacccagcataaaaaattgcag aagtcaa | | | | | |
| **HAF** | chr3 | 54907512 | no SNP | agtgacctactagacaaagagt tcaaacaaaatttcatgaggatg ctcacagatatgggaagaaga attgatgaacacagtgagcaca tcagcaaagaactggaagatat aaaaaaaattagaaatgaggaa tagaatactggaaatgagaaatt cactagagggactcaacagca gaatagaggaagcagaagaac ggatcagcgagctagatgaaa gactagaggaaatcacccaag cagaacacaaaagagaaaaaa gaattagacagaatgagaaca gtctaagggaactctggga[c/t ]aatatcaagcatgctaacatttg gattataggtgtcccagaagga gaagagagagacaaaggggc ataaaatttatttgtagaaataata gacgaaaattttcctaacatgag gaaggaaacagacatccaagtt caggaagcacagagagctcca aacaagataagcccaaagagg cccacaccaagacatattataat taaaatgtccaaaattaaaaaca aagagagaaccctaaaagcag caagagaaaggccacaagtga catataaagggaagcccatcag gctatcagcagac | chr3:549 07512.f | aacggatca gcgagctag at | chr3:54 907512. r | atatgtctt ggtgtgg gcct | 303 |
| **HAF** | chr1 4 | 92870205 | no SNP | cccttggagtttcctcagtgaca tgggggagaggagcttcttttgt cattcataacaagcccatttcaa ccctacctgatttatgcaatgag gtgactccaggtggggcccct ggatggcttcaggatggggac cggtggccaggggatccaacc atgtgatcagagggttggaactt tcagccccaccccagaccttca gggagggagaggggctgaag gttgagttcattcacccatggcc agtgatttagttgttcatgtctaa gaatggaacctccataaaaccc ctaaatgatgg[g/a]gtttgga gagcctccttattgggtgggtgt cacaccccagaccccacaggg acagaagctcctacactaggga ccgttccagacctcgcctcatgt acctcttcatctgactgttcatgt gtatcctttacaataaaccagta aacttaagtgttttcctgggttct gtgatacattctagcaaatcacg gaagctgagggaggggtcatg | chr14:92 870205.f | ccatgtgatc agagggttg g | chr14:9 2870205 .r | gccaccc agacttat gacca | 404 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | gggacccctgatttagagccag<br>ttggtcataagtctgggtggcaa<br>cctgggactcaggactgccctc<br>tgaactgagggctgctgtggga | | | | | |
| **HAF** | chr1 | 143424677 | no SNP | attgcctgtttttgtggttgtgtgtt<br>tgggtttttgataatttttttatct<br>ctatttatggtcattgctttcccac<br>ttaaataagtccctttagcatttct<br>tgtagaactggtttcttagtgata<br>aactcctttaatttttgcttgtttgg<br>gaagctctttatctctccttccatt<br>ctaaatgacagccttgatggata<br>gagtattcttggttgtaggttttttt<br>ccttttagcactttaaatatgtcgt<br>gccattctcttctcgcctgtagg<br>gtctcgactgagaagtctgctg<br>acag[c/a]ctgatgggcttccc<br>tttatatgtcacttgtggcctttct<br>cttgctgcttttaggattctctctttt<br>gtctttaattttagacattttgatta<br>taatatgtcttgatgtgggcctct<br>ttgggcttctcttgtttggagctct<br>ctatgcttcctgaacttggatgtc<br>tgtttccttcctcaggttaggaaa<br>attttcctctattatttcgacaaat<br>aaattttctgccccttttgtctctct<br>catctccttctgggacccctata<br>atccaaatgttagcacgcttgat<br>attgtcccagagttcc | chr1:143<br>424677.f | tggtcattgct<br>ttcccactt | chr1:14<br>3424677<br>.r | gggacaa<br>tatcaagc<br>gtgct | 535 |

# APPENDIX D

## IGV AND SANGER VISUALS

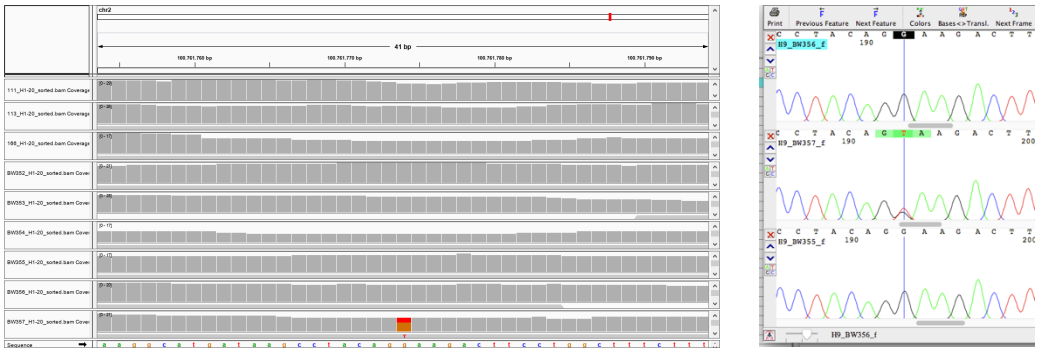

**Figure A.D.1:** IGV screenshot and Sanger result of a 'de novo' candidate (chr2:100761774) which was validated truly in BW-357 in the top and left respectively. None of the individuals except BW-357 had the mutation.

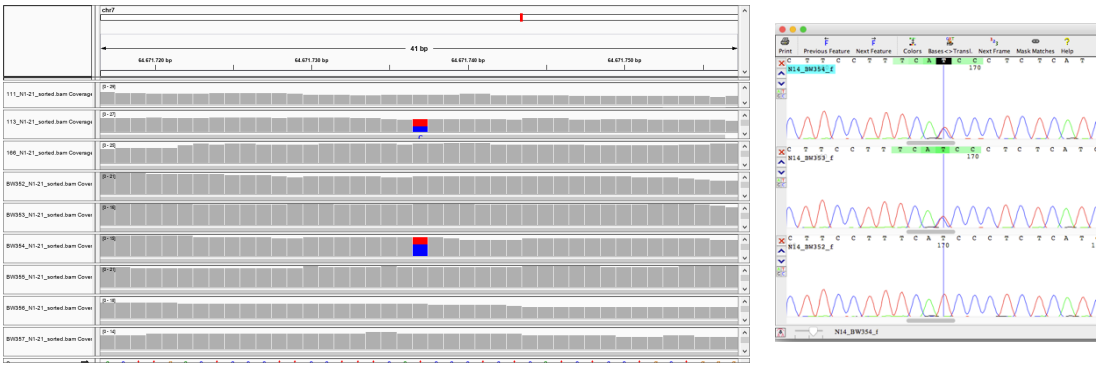

**Figure A.D.2:** IGV screenshot and Sanger result of a ‘de novo’ candidate (chr7:64671737) which was also detected in one of the parents of BW-354 in the top and left respectively. However, it seemed like parents do not carry the mutation in the IGV screenshot, Sanger result detected the mutation in BW-353 (Noriker mother).

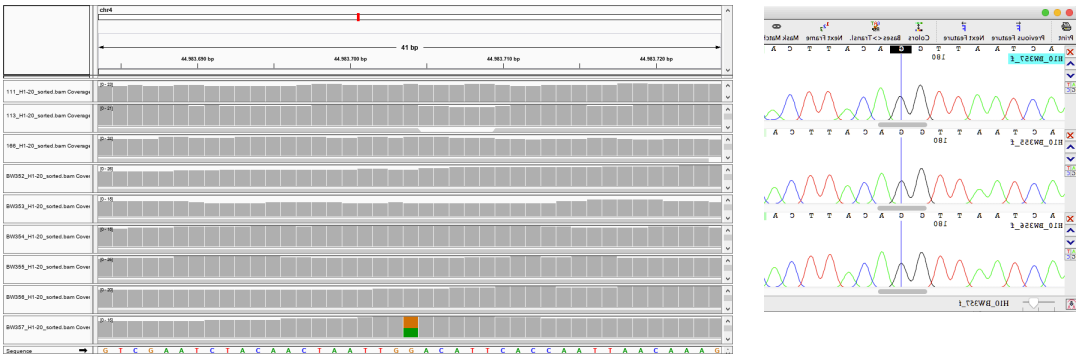

**Figure A.D.3:** IGV screenshot and Sanger result of a ‘de novo’ candidate (chr4:44983704) that called as no SNP depending on the validation result in BW-357. However, the candidate seems like a ‘de novo’ mutation, Sanger sequencing did not detect any variant at the position.

# APPENDIX E

## LOGARITHMIC VERSION OF PSMC

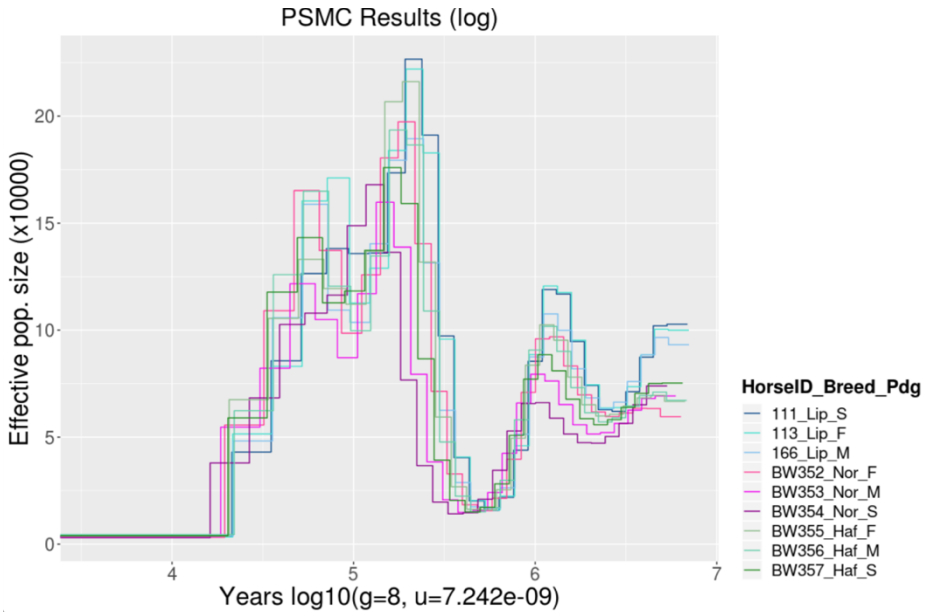


**Figure A.E.1:** Logarithmic version of PSMC graph. All horses showed similar patterns in the graph, but 111 and 113 had a larger effective population size than the other trios during the timeline.