Applying massive parallel sequencing technologies to unravel the genomic complexity in domestic species

Agnese Viļuma

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2017

Acta Universitatis agriculturae Sueciae
2017:93

Cover: Designed by the author

ISSN 1652-6880
ISBN (print version) 978-91-7760-076-3
ISBN (electronic version) 978-91-7760-077-0

Print: SLU Service/Repro, Uppsala 2017

Applying massive parallel sequencing technologies to unravel the genomic complexity in domestic species

Abstract

Massive parallel sequencing technologies have made a remarkable contribution in understanding the genomic complexity in human and other mammalian species. This thesis includes four studies illustrating the technology change from short-read to long-read massive parallel sequencing and the corresponding layers of genomic complexity resolved in dog and horse.

In the first study, the Ion Proton sequencing platform was evaluated for whole-genome re-sequencing of four dog genomes. On average, 80 % of the genome and 77 % of the exome had at least 4-fold coverage. The obtained genotypes showed 90 % concordance with those using the CanineHD BeadChip. Next, whole-genome re-sequencing of a canine family trio was used as a proof-of-concept to map an autosomal recessive disorder termed hereditary footpad hyperkeratosis in the Kromfohrländer breed. This resulted in the identification and association of a single nucleotide variant in the *FAM83G* gene.

In the third study, we generated the first long-read assembly of the horse MHC class II region (1.2 Mb) by assembling eight bacterial artificial chromosome clones sequenced with single molecule, real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio). Detailed annotation of this assembly revealed 23 functional genes and 12 pseudogenes. In comparison to other mammals, the horse MHC showed an increased number of MHC class II genes in different relative positions and with different directionality.

In the fourth study, "sequence self-similarity" of the equine MHC class II region was investigated to identify paralogous segments and to reconstruct the evolutionary history of duplication events. The results indicated a major expansion of class II loci 55-70 million years ago, coinciding with mammalian radiation. We showed that the high genomic structural diversity and plasticity of the mammalian MHC region may be a result of recurring segmental duplication events mediated either by unequal crossing-over between misaligned genomic regions or, possibly, by alternative transposition of long interspersed nuclear elements.

Keywords: whole genome re-sequencing, trio sequencing, dog, hereditary footpad hyperkeratosis, horse, MHC, annotation, gene families, unequal crossing-over, transposable elements

Author's address: Agnese Viļuma, SLU, Department of Animal Breeding and Genetics,
P.O. Box 7023, 750 07 Uppsala, Sweden

Massiv parallellsekvensering av domesticerade arter för kartläggning olika nivåer av genetisk komplexitet

Abstrakt

Olika DNA-sekvenseringstekniker som gemensamt kallas ”Massive parallel sequencing technologies” har bidragit till ökad förståelse rörande genomisk komplexitet hos människan och andra däggdjursarter. Avhandlingen innehåller fyra studier som illustrerar teknikförändringen från tidigare versioner av tekniken som endast gav korta sekvenser till nu tillgängliga metoder som kan producera långa sekvenser vilka kan användas för att kartlägga olika nivåer av genomisk komplexitet hos hund och häst.

I den första studien utvärderades sekvenseringsplattformen IonProton för helgenom-sekvensering av fyra hundar. I genomsnitt uppvisade 80% av genomet och 77% av exomet en 4-faldig täckningsgrad (4X). Genotyperna uppvisade 90% överensstämmelse med resultat från CanineHD BeadChip. I en efterföljande sjukdomsgenetisk studie i syfte att identifiera den orsakande mutationen för en autosomal recessivt nedärvd hundsjukdom som kallas digital hyperkeratos. Helgenom-sekvensering av en familje-trio bestående av friska föräldrar och deras sjuka valp utfördes och resulterade i identifiering av en basparssubstitution i genen *FAM83G* som orsakar ett aminosyrautbyte från arginin till prolin.

I den tredje studien användes ny sekvenseringsteknik från Pacific Biosciences (PacBio) för att producera långa sekvenser med så kallad ”single molecule, real-time (SMRT) sequencing” av hästens Major Histocompatibility Complex (MHC) klass II-region (1.2 Mb). För ändamålet sekvenserades åtta BAC-kloner. Detaljerad sekvensannotering av gener i regionen resulterade i att 23 funktionella gener och 12 pseudogener identifierades. I jämförelse med andra däggdjur uppvisade hästens MHC ett ökat antal paraloga MHC klass II-gener vilka var positionerade i olika inbördes ordning och riktning.

I den fjärde studien undersöktes sekvensen av MHC-klass II-regionen mot sig själv för att identifiera paraloga segment och för att definiera duplikationernas storlek och dess evolutionära historia. Resultaten indikerade en större expansion av klass II-loci för 55-70 miljoner år sedan vilket sammanfaller med tidpunkten för däggdjurens artexpansion. Vi visade att den omfattande strukturella diversiteten och plasticiteten hos däggdjurens MHC-klass II region kan vara ett resultat av flertaliga segmentala duplikationer medierade av icke-reciproka rekombinationshändelser eller alternativt via s.k. ”alternative transposition” av LINE retrotransposoner.

Nyckelord: genomsekvensering, trio-sekvensering, hund, digital hyperkeratos, häst, MHC, annotering, genfamiljer, rekombination, transposoner

Författarens adress: Agnese Viļuma, SLU, Institutionen för husdjursgenetik,
P.O. Box 7023, 750 07 Uppsala, Sweden

Dedication

In memory of my most loyal friend

Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.
Albert Einstein

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Viļuma A.*, Sayyab S.*, Mikko S., Andersson G. and Bergström T.F. (2015). Evaluation of whole-genome sequencing of four Chinese crested dogs for variant detection using the ion proton system. *Canine Genetics and Epidemiology*, 2:16.

II Sayyab S.*, Viļuma A.*, Bergvall K., Brunberg E., Jagannathan V., Leeb T., Andersson G. and Bergström T.F. (2016). Whole-Genome Sequencing of a Canine Family Trio Reveals a *FAM83G* Variant Associated with Hereditary Footpad Hyperkeratosis. *G3*, Jan 8;6 (3), pp. 521-527.

III Viļuma A., Mikko S., Hahn D., Skow L., Andersson G. and Bergström T.F. (2017). Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology. *Scientific Reports*, Mar 31;7:45518.

IV Viļuma A., Mikko S., Andersson G. and Bergström T.F. Evolution of the horse MHC class II region by gene duplication. (Manuscript)

Papers I-III are reproduced with the permission of the publishers.

* These authors contributed equally

Abbreviations

BAC	Bacterial Artificial Chromosome
dNTP	Deoxynucleotide triphosphates
ERE	Equine repetitive elements
Gb	Giga base
GWAS	Genome-wide association studies
HFH	Hereditary footpad hyperkeratosis
INDEL	Insertion or deletion
kb	Kilo base
LINE	Long interspersed nuclear element
Mb	Mega base
MHC	The Major Histocompatibility Complex
mRNA	Messenger Ribonucleic acid
MYA	Million years ago
OMIA	Online Mendelian Inheritance in Animals
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
rDNA	Ribosomal Deoxyribonucleic acid
SINE	Short interspersed nuclear element
SMRT	single molecule, real-time
SNP	single nucleotide polymorphism
SNV	Single nucleotide variant
tDNA	Transfer Deoxyribonucleic acid
TE	Transposable element
UTR	Untranslated region
WES	Whole exome-sequencing
WGS	Whole-genome sequencing

1 Introduction

Life on Earth is known for outstanding phenotypic diversity and countless adaptations to a vast variety of environmental conditions that have ensured survival of living organisms for billions of years. This extraordinary diversity has amazed and puzzled many great minds for centuries. In the 18th century Carl Linnaeus named and classified living organisms into species creating taxonomy that has been widely used and appreciated in science. About a hundred years later, in 1859, biologist Charles Darwin introduced the theory of evolution to general society by publishing his work "On the Origin of Species" (Darwin, 1859). During the same time, understanding the ability of passing phenotypic traits from one generation to the next in pea plants *Pisum sativum* was the passion of Gregor Mendel. He established the rules of Mendelian inheritance and introduced the principles of "recessive" and "dominant" phenotypic traits, terms that are still being used in genetics to characterise the inheritance pattern of "Mendelian traits" - phenotypes influenced by a single gene. In the early 20th century, Thomas Hunt Morgan took this knowledge to the next level by using the fruit fly *Drosophila melanogaster* to scientifically prove the role of chromosomes and crossover in heredity. These, and many other scientists of that time, remarkably contributed to the understanding of genetics and evolution based solely on phenotypic observations and considering genes as invisible factors of inheritance.

Discovering DNA as hereditary material (Avery et al., 1944) and understanding the structure of the DNA helix (Watson and Crick, 1953) were important landmarks for the field of molecular genetics and brought a new twist to future research. From then, great efforts were focused on accessing and understanding the information carried by DNA. Just a few decades later, Fred Sanger published a DNA sequencing method using chain-terminating inhibitors (Sanger et al., 1977). This idea was later developed in technology that was used to pioneer the knowledge of the genetic code, commonly known as Sanger-sequencing. In 1996, the first eukaryotic genome 12 Mb in size, *Saccharomyces* *cerevisiae* known as baker's yeast, was fully sequenced (Goffeau et al., 1996). Soon after, the first draft of a human reference genome was published (Venter et al., 2001, Lander et al., 2001). This effort cost billions of USD and took ten years of laborious work from two independent groups using different approaches to solve the human genome puzzle. While the human genome sequencing consortium took advantage of certainty provided by linkage maps and shotgun sequencing of Bacterial Artificial Chromosome (BAC) clones (Lander et al., 2001), Craig Venter's initiative pursued the idea of whole-genome shotgun sequencing (Venter et al., 2001). These two projects highlighted the challenges of any genome assembly, showing that the final result is a compromise between cost and time effectiveness, and quality of the assembly. During the following decade, by a hybrid approach of using BAC/BAC-end and whole-genome shotgun sequencing the most of the domestic animal species acquired a reference genome assembly.

After publishing the human genome sequence, single nucleotide polymorphisms (SNPs) became "the bread-and-butter of DNA sequence variation" (Stoneking, 2001). Consequently, large-scale SNP identification with DNA chips (Wang et al., 1998) contributed to gathering large amounts of SNP variation in human and domestic animals and a new analysis method, called genome-wide association (Ozaki et al., 2002) allowed to link this genetic information to observed phenotypic variation.

The launching of massive parallel sequencing in 2005 was another vital landmark that further revolutionized the field of molecular genetics. The reduced cost and increased efficiency has enabled scientists to detect and link the genetic variation in domestic species to productivity, health and exterior phenotypes. Furthermore, with the continuous increase in massive parallel sequencing read length, we are now on the brink of exploring the "hidden portion" of the structural variation "iceberg" and understanding genomic plasticity and the mechanisms behind it.

2 Background

2.1 Genomic complexity in domestic species

In his monograph “The variation of Animals and Plants under Domestication” Darwin described a variety of artificial selection examples in domestic species. Artificial selection takes advantage of natural genomic variation present in domestic animals and resembles natural selection, except that it is driven by mankind instead of natural processes (Driscoll et al., 2009). However, the principle remains the same – the population moves towards the best possible phenotypes for the existing environment, in this case, human needs.

The role of genetic factors in the adaptive change of phenotypes has been intensively studied and a common consensus understanding has been reached that the genomic variation provides grounds for phenotypic variation, which is essential for adaptation and survival. In other words, environment selects for the most advantageous genotypes through phenotypes which, in fact, are controlled by genotypes (Fusco and Minelli, 2010). The genomic variation pool of domestic species consists of the variation that was present in wild ancestors at the time of domestication and private variation that has accumulated since domestication. Humans have taken a great advantage of this limited genomic variation to drive the formation of diverse breeds that drastically differ based on their appearance, productivity and disease resistance. A classic example of this phenomenon is phenotypic variation of dog breeds, that have been created by humans despite the limited ancestral gene pool (Driscoll et al., 2009). Due to vast phenotypic and limited genomic variation, domestic species have been excellent models for unravelling the mammalian genomic complexity. For simplicity, genomic complexity will be further dissected into two major components, where the first component describes genomic plasticity and mechanisms behind it and the second component describes how genotypes manifests as phenotypes.

2.1.1 Genomic plasticity

The genomic "blueprint" of a living organism is not a static, but rather a dynamic entity. The nucleotide sequence of a deoxyribonucleic acid (DNA) molecule is prone to changes (mutations) of different magnitude and the ability to alter DNA sequence is known as genomic plasticity. Mutations can occur in any type of cells. However, for a mutation to manifest as a phenotype in the next generation, it must occur in germline. There are several mechanisms for DNA alterations to occur. The known mechanisms include DNA replication errors, polymerase slippage, recombination, unequal crossing-over during cell division and integration of foreign DNA.

Single nucleotide variants (SNVs), also referred to as SNPs if the minor allele is present in a frequency of >1% in the population, are the most studied type of genomic variation. Genetic variation that is caused by the insertion or deletion of a single base pair or base pair sequence is denoted as INDELs, as it is often not possible to distinguish between the two events without the knowledge of the ancestral sequence. INDELs can be of variable size, spanning from 1 bp to several Mb. However, large INDELs and inversions are often denoted as structural variation.

A large proportion of eukaryotic genomes corresponds to repeated DNA sequences. Repetitive DNA contributes to the sequence length variation of different magnitudes and can be divided into a) tandem repeats, *i.e.* tandem paralogs, satellite DNA (satellites, minisatellites and microsatellites) and rDNA, and b) dispersed repeats, *i.e.* paralogs, DNA and retrotransposons, tDNAs and retrogenes (Richard et al., 2008). The term copy number variation (CNV) is used to illustrate the variable number of repetitive DNA segments (>1 kb) between individuals.

2.1.2 From Genotype to Phenotype

Regardless of phenotypic differences, when compared to other groups of life, mammalian genomes show relatively smaller variation of the genome size than for example, frogs, salamanders, crustaceans, flowering plants and many other species (Gregory, 2005). In addition, compared to the initial estimate, the number of protein-coding genes in the human genome has decreased to around

19 000 (Ezkurdia et al., 2014) which is in fact close to that observed in a genome of a roundworm *Caenorhabditis elegans* (Hillier et al., 2005). Furthermore, the majority of the genes in the human genome are present also in other mammalian species and share a high degree of nucleotide sequence similarity (Lindblad-Toh et al., 2011). Most of eukaryotic genomic complexity relies on the fact that a single gene can be transcribed into a variety of mRNA sequences due to the existing exon-intron structure and on the ability to alter gene expression level, transcription start sites and polyadenylation.

Mutations can occur at any genomic loci, but just a proportion of those can further alter the phenotype. The most straight-forward way for a mutation to alter the phenotype is to coincide with the protein coding part of a gene. A SNV present in an open reading frame of a protein-coding gene can result in a synonymous (or silent) mutation where the amino acid content of the protein remains un-altered or it can be non-synonymous and change the amino acid content (missense mutation) or prematurely terminate translation of the protein (nonsense mutation). The effect of INDELs usually depends on the size of the mutation. Single and double nucleotide INDELs are typically responsible for non-sense mutations, while triplet INDELs usually produce milder consequences by deletion or inserting a single amino acid residue. While non-synonymous SNVs and INDELs are more likely to influence the phenotype, it has been shown that synonymous mutations can have an effect on translational level by influencing the structure of the protein (Kimchi-Sarfaty et al., 2007). Incorporation of large transposable elements (TEs) in the intronic region or single nucleotide splice site mutations can cause skipping of exons (alternative splicing) and result in new transcripts. Large structural variation or transposition that occurs in coding regions may completely disrupt the function of the gene.

Gene expression can be influenced in both directions, either promoting or decreasing the mRNA production, otherwise known as transcription. TEs are common factors that influence gene expression (Faulkner et al., 2009). To name a few examples, TEs by altering gene expression were shown to influence the colour of maize kernels (McClintock, 1951) or decrease of *MSTN* gene expression levels (Santagostino et al., 2015). The most striking of all examples being the TE involvement in evolution of mammalian pregnancy (Lynch et al., 2015). The gene expression can also be altered by segmental duplication of the entire gene. For example, adaptation to a starch rich diet in dogs was associated with increased copy number of the Amylase gene (Axelsson et al., 2013). Furthermore, non-coding RNAs add an extra layer to mammalian genomic complexity and have been shown to have an impact on gene expression (Carninci, 2008) and transcriptional start sites (Taft et al., 2009).

2.2 Massive parallel sequencing technology

The cost efficient and relatively simple techniques of analysing genomic sequence of living organisms position molecular genetics as an important part of medicine, biology, ecology, evolution, agriculture, conservation, forestry, food safety and many other fields. The first massive parallel sequencing technology became commercially available in 2005 and during the last decade it has considerably evolved in terms of accuracy, throughput, read length and speed. Currently, there are various platforms based on different technological principles, reviewed in (Heather and Chain, 2016) and (Goodwin et al., 2016). Existing technologies can be divided in two groups – short-read and long-read sequencing based on the technological limits on the generated read size (Figure1).

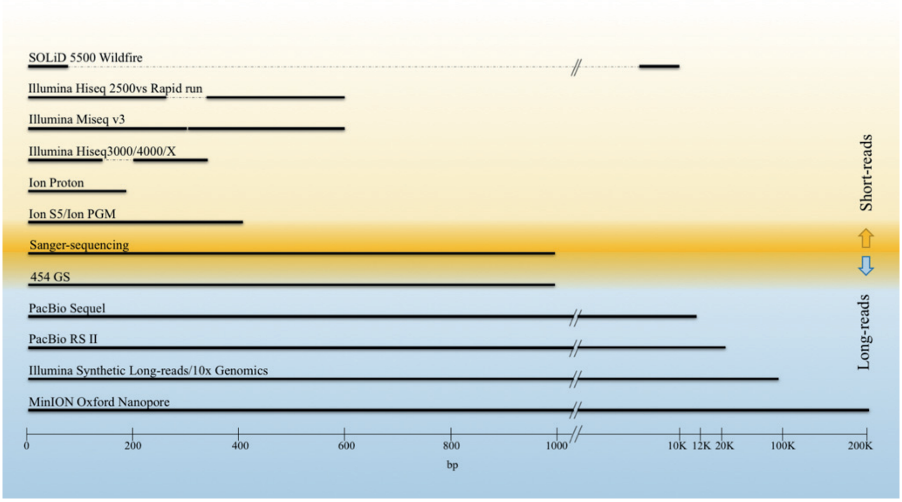

Figure 1. Schematic overview illustrating the read-length variation of massive parallel sequencing platforms based on read length summary from (Goodwin et al., 2016).

2.2.1 Short-read sequencing technologies

Pyrosequencing, the pioneering technology of massive parallel sequencing, used a recently discovered luminescent method that allowed measurement of pyrophosphate production. During the pyrosequencing process the template DNA is washed with each nucleotide in turn and a pyrophosphate signal is emitted at each successful incorporation (Heather and Chain, 2016). While the principle of pyrosequencing was introduced already in 1993 (Nyren et al., 1993), it took almost ten years until the first machines, capable of the mass parallelisation of sequencing reactions and production of 25 million bases in one four-hour run, were introduced to the market (Margulies et al., 2005). Soon after, two similar technologies emerged, Illumina (previously known as SOLEXA) and SOLiD (sequencing by oligonucleotide ligation and detection).

Illumina sequencing, so far the most successful short-read technology, is based on the sequencing-by-synthesis principle, where differently fluorescent deoxynucleotide triphosphates (dNTPs) are detected after their incorporation (Heather and Chain, 2016). The advantages of Illumina sequencing are the low error rate and the ability to sequence both ends of a larger DNA fragment (up to 600 bp) due to a solid phase "bridge-amplification" process (Heather and Chain, 2016). Illumina technology became the first choice for many researchers and over time it has been further developed from sequencing 30 bp paired-end reads to currently five different throughput sequencing platforms that can sequence paired-end reads of 125, 150 and 300 bp (Goodwin et al., 2016, Bentley, 2006).

An alternative non-optical sequencing technology, Ion Torrent, based on the previously described pyrosequencing principle was launched in 2010. This technology measures a pH change caused by the release of protons during polymerisation (Rothberg et al., 2011). Due to lower throughput, this technology primarily gained its popularity mainly for sequencing smaller genomes, but later was developed into the Ion Proton system with higher throughput capacity suitable for whole genome sequencing of mammalian genomes.

2.2.2 Long-read single molecule sequencing technologies

Currently there are two single molecule sequencing technologies, based on cardinally different principles. Both technologies are not only capable of generating impressively long reads, but can also bypass GC-coverage bias common to polymerase chain reaction (PCR) based short-read sequencing. Single Molecule, Real Time (SMRT) technology from Pacific Biosciences has already established its place in the market and has been used to improve several mammalian reference genomes and to solve complex repetitive regions (Vij et al., 2016, Hans et al., 2017, Chaisson et al., 2015). SMRT sequencing takes place in tiny wells called zero-mode waveguides, where single fluorophore molecules are visualized as a new nucleotide is incorporated during template DNA polymerisation (Levene et al., 2003). In addition to sequence data, this technology can also provide DNA methylation data as nucleotide incorporation happens real time and incorporation of modified nucleotides takes more time (Flusberg et al., 2010). Even though SMRT sequencing suffers from high error rate (up to 13% for single pass reads), most of the errors are of random nature and are easily removed to < 1% by repeated sequencing of the same molecule (Chin et al., 2013, Goodwin et al., 2016)

MinION is the first nanopore sequencer offered by Oxford Nanopore Technologies (ONT). Nanopore sequencing is conducted in a pocket-sized device by threading an intact DNA strand through the nanopore and determining bases one at a time, based on current amplitude (Magi et al., 2017). While retaining slightly higher error rates of approximately 5 to 8 % (Jain et al., 2015) this technology is capable of generating ultra long reads up to 1 Mb (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/). This technology is especially valued for new bacterial genome references and targeted amplicon sequencing. Nevertheless, the first ultra long-read assembly of the human genome was recently produced (Jain et al., 2017).

2.3 Bioinformatics

The research field of bioinformatics was introduced by two Dutch scientists, Ben Hesper and Paulien Hogeweg in 1970 to emphasise the importance of studying the information flow in living systems (Hogeweg, 2011). Nowadays, a massive parallel sequencing run can generate up to 900 Gb (Goodwin et al., 2016) of sequence data from genomic DNA, transcriptomes, interactions with proteins or meta-genomes, creating a major challenge for field of bioinformatics to evolve together with technological achievements and supply methods to study complex dynamic systems.

The raw data of a sequencing run is most commonly obtained in the form of a simple text file with read identifiers, sequence of nucleotides and quality scores for each read. As the number of generated reads can be measured in billions, the existence of appropriate bioinformatics tools is vital for further analysis and visualization of the obtained information. To accommodate this scientific need, a plethora of publicly available tools have been developed and user support forums formed. In addition, some of the sequencing technologies provide their own alignment and/or assembly tools. For example, TMAP aligner is designed to cope with INDEL errors of Ion Proton data and HGAP assembler (Chin et al., 2013) can cope with long reads with high error rates. There are several steps of basic bioinformatics analysis that need to be followed before one can proceed with the further analysis of interest. Accordingly, the quality of the final conclusions will be highly dependent on the accuracy of these initial steps.

2.3.1 Genome re-sequencing

Genome re-sequencing is a straight-forward sequence read analysis to obtain the genomic variation necessary for phenotypic trait mapping, evolutionary analysis, selection sweep detection etc. However, it requires an existing knowledge of a reference genome from the investigated or, in worst-case scenario, closely related species. By comparing each sequence read to the reference genome with a suitable alignment tool (review of different algorithms by (Schbath et al., 2012, Caboche et al., 2014), the most likely location (the origin) of the read is found. The choice of the alignment algorithm depends on the sequencing platform that has been used to generate the data. For example, algorithms that perform well with mapping of short and low error rate Illumina reads, perform less efficiently as the length and sequencing error rate increases, as in case of Ion Proton and PacBio data. There are two main difficulties in read alignment, i) unmappable reads (mapping quality 0), and ii) reads with poor mapping quality (potentially incorrect mapping). Both issues are caused by repetitive sequences and segmental duplications (Li and Freudenberg, 2014). As suggested by (Li and Freudenberg, 2014), the mappability of reads increases together with the increase of read length.

After all possible reads have been mapped to the reference genome, a variety of tools can be used to investigate genomic differences between the reference genome and the aligned reads. The most common types of difference that can be detected with short-read sequencing are SNVs and small INDELs. With the advent of long-read sequencing a structural variation in terms of larger INDELs, inversions and translocations can be assessed. The success of detecting variants relies on the number of reads that are covering each base of the reference genome (depth of coverage) and a performance of the bioinformatics tool that is chosen for analysis (Hwang et al., 2015). When adjusting the tool performance, the usual compromise is between sensitivity and specificity that can be achieved. On one hand, the stricter the quality parameters are set, the less of the called variants will pass the quality control. On the other hand, the less strict, the more errors will be called as true variants. In cases when a comprehensive annotation and enrichment assay of an exome is available, it is possible to sequence only the coding portion of the genome, reducing the costs per experiment and increasing depth of coverage per sample.

2.3.2 *De novo* assembly

In cases when the reference genome sequence is not available or the genome of interest needs to be investigated independently, a *de novo* assembly of the raw reads is required. During the assembly process, all of the generated reads are compared to each other to find the overlapping sequences that would further allow to join them in a single continuous sequence spanning the entire chromosome, in the best case scenario. Based on the concept of sequence assembly described by R. Staden (Staden, 1979), if the overlap is of sufficient length to distinguish it from being a repetitive sequence, the two compared sequences must be contiguous. The data from contiguous reads can then be further joined to form a longer continuous sequence, called a contig. Nowadays, there is a broad choice of assembly algorithms, listed in (Sohn and Nam, 2016). Repetitive DNA is the biggest challenge of the assembly process, that often can not be solved with short-read sequencing and the full length continuous assembly of the entire chromosome is not possible. Instead, a set of contigs of different length is generated. Consecutively, the next step of the assembly process is to determine the correct order of contigs and join them in an ordered manner forming scaffolds. A common approach of scaffolding is to use paired-end (Illumina) libraries with larger insert size, mate-pair (SOLiD, Illumina) or long-read sequencing (PacBio, Nanopore) in order to span the difficult repetitive regions and join the individual contigs (Hunt et al., 2014). Even though these methods are sufficient to resolve considerable portion of repetitive complexity of the genome, some extra steps are required to achieve chromosome scale scaffolds. Approaches like BAC and fosmid clone sequencing, optical mapping, linked read sequencing, synthetic long reads and genome-wide chromatin interaction data have been proposed to address this issue (Ghurye et al., 2017). *De novo* assembly can also be attributed towards the assembly of the transcriptome, where instead of sequencing and assembling the genomic DNA, an entire set of mRNA extracted from a sample is sequenced and assembled. The assembly of a transcriptome has its own specificities and tools reviewed in detail by (Martin and Wang, 2011).

2.3.3 Sequence Annotation

Often, scientific interest is focused on genomic variants that result in phenotypic alterations. Thus, the large amount of detected genomic variants need to be sorted into variants that putatively alter the phenotype, like variation that is located in the protein-coding regions or regulatory elements, and those that most likely do not. If the species under investigation possess a good annotation and is part of the 29 mammals' comparison (Lindblad-Toh et al., 2011), the task of finding this variation is fairly easy. Tools like ANNOVAR (Wang et al., 2010) extract the variation that is present in genomic regions defined by the existing annotation and classifies the type of mutation, whether it is altering the amino acid, prematurely terminating translation or staying silent.

Protein-coding genes possess certain structural characteristics, like transcription start sites, 5'- and 3'-UTRs, conserved start and stop codons, splice sites, poly A signals etc. These typical features make it possible to screen the genomic sequence with annotation tools (Wang et al., 2004) to define the potential location of the open reading frames encoding protein sequences. However, this approach is often imprecise (Reese et al., 2000) due to pseudogenes that are missing typical features or, simply, due to the natural variation and exceptions of different organisms. The best evidence of an existing gene is an aligned mRNA sequence. The current public databases contain a considerable amount of evidence, like short expressed sequence tags, Sanger-sequenced mRNAs and protein sequences that have been collected for many years and can be used as valuable extrinsic evidence to aid in better annotation. Massive parallel sequencing provides an affordable way to sequence the entire transcriptome of a sample. The annotation process begins with mapping the mRNA evidence to the reference genome. Algorithms for mRNA mapping are different from those of genomic read mapping due to exon-intron structure of the gene. Further, there are tools, like for example AUGUSTUS (Stanke and Waack, 2003), that is capable of incorporating extrinsic evidence in the gene prediction models.

2.3.4 Quality control

While doing large scale analysis it is not possible to check every detail "by eye" and it is of great importance to perform a proper quality check at every analysis step, starting from raw read quality control until variant discovery and annotation. Universal quality score system has been invented to facilitate the quality control of large scale genomic analysis.

Quality of the base calling. The accuracy scores of each base in a read are based on an already known system of PHRED error-rate prediction (Richterich, 1998), similar to Sanger-sequencing quality scores. The probability of error for each base call has been estimated as a function of certain parameters calculated from trace data (Ewing and Green, 1998). The base call accuracy of 99.9% is achieved with a PHRED accuracy score of 30 (Q30), which means that incorrect base call probability is 1 in 1,000. This score is accepted as benchmark for the quality of a base call, and reads of lower quality base calls can be either eliminated from analysis completely or trimmed, if the low quality occur at the ends of the read.

Quality of the read mapping. For each mapped read, the mapping quality is calculated based on the similarity of the read to the mapped segment of DNA and the chance that it could have originated from an alternative location. If there are several locations where the particular read would align with the same score, the read is assigned mapping quality of 0. Mapping quality threshold can be used, for example, when calling SNVs (Li et al., 2008).

Quality of the variant calling. Each detected variant, whether it is a SNV or INDEL, has a quality value that can be calculated using different methods (Nielsen et al., 2011). These values can be used to filter away dubious variants, as they reflect the uncertainties due to coverage, sequencing and read mapping quality (Nielsen et al., 2011). The observed variant can be classified as a transition, when the nucleotides of similar structure are exchanged, or a transversion, when nucleotides of different structure are exchanged. If a transition-transversion ratio for the current species is known, it can be used as a guide for assessing the reliability of the obtained results (DePristo et al., 2011).

Quality of the assembly. The quality of assembly is difficult to measure and it is usually described by the size and accuracy of the contigs and scaffolds (Miller et al., 2010) The most common way of assessing the quality of the assembly is by calculating the N50 statistic. This statistic describes the contiguity of the assembly, where 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than the N50 value (Miller et al., 2010).

2.4 Inherited disease mapping in domestic animals

Understanding the underlying genetic causes of disease phenotypes is an important contribution to better welfare of domestic animals and decreased economic losses for owners. Inherited diseases can be characterized in two categories - Mendelian disorders, where the disease phenotype is caused by a mutation in a single gene, and complex disorders, where the disease phenotype is caused by combination of mutations in several genes and the environment. Mendelian disorders, while rare, are collectively common and have a major impact on animal health. Discovery of causative mutations allows for the development of gene tests to identify carriers of the causative genetic variants and to lower the frequency of disease-causing mutations in a population of domestic animals.

Since the rediscovery of Mendelian inheritance, scientists have been noting traits that segregate according to Mendelian laws and after Alfred Sturtevant laid out the logic of genetic-linkage mapping in 1913 it became a key tool for linking Mendelian disease phenotypes to a certain chromosomal location (Nicholas, 2005). For example, economic loss due to Mendelian disorder, termed malignant hyperthermia, in pigs encouraged the development of the porcine linkage map that, in combination with comparative genomics, allowed the pinpointing of the causative mutation (Fujii et al., 1991). With development of DNA chips, genome-wide association studies (GWAS) of SNP allele frequencies in case-control settings, became a successful tool for discovering common variants causing many Mendelian disorders and gave some insight into the genomic architecture of complex disease, reviewed by (Goddard and Hayes, 2009, Day-Williams and Zeggini, 2011). The use of massive parallel sequencing technologies has made it possible to identify rare single nucleotide variants and structural variation allowing previously unattainable association of these variants with phenotypes of interest (Day-Williams and Zeggini, 2011). Recently, a clinical case has been published where long-read genome sequencing of a diagnosed patient identified causal structural variation in a Mendelian disease, that was previously overlooked with short-read sequencing (Merker et al., 2017).

Information of Mendelian traits and inherited disorders are catalogued in the Online Mendelian Inheritance in Animals (OMIA) database that currently reports 3363 traits and disorders of which 1337 are of Mendelian inheritance. In 733 of Mendelian traits and disorders the causative mutation is already known (http://omia.angis.org.au/home/, accessed on 26th of September 2017).

2.5 Major Histocompatibility Complex

The Major Histocompatibility Complex (MHC) is a gene dense region and over 40 % of the expressed MHC genes have a function in the immune system (The MHC sequencing consortium, 1999). Among them are some of the most allele rich gene families, that encodes for antigen presenting molecules and are present in all the jawed vertebrates (Flajnik and Kasahara, 2001). Originally, this region was discovered while studying the rejection of allogeneic tumour transplants in mice (Gorer et al., 1948) and was initially designated as the Histocompatibility 2 (H2) locus. This historical designation is still retained in nomenclature of mouse MHC genes.

Defining the structure, function, and diversity of the MHC region is key to understanding the immune response in vertebrate species (Fraser and Bailey, 1998). Based upon the functional differences, genes in this region are traditionally divided into MHC Class I, II and III. The classical class II cell surface glycoproteins bind and present exogenous peptides to T-lymphocytes, initiating the immune response. These glycoprotein molecules are heterodimers formed by α- and β-chain peptide encoded by two class II genes. Thus, classical class II molecules DR, DQ and DP are encoded by *DRA* and *DRB*, *DQA* and *DQB*, *DPA* and *DPB* genes, respectively. The most polymorphic sites of these genes are the antigen recognition sites, or more specifically - the binding cleft and this polymorphism is driven by the need to maximize peptide binding diversity (Hughes et al., 1990, Brown et al., 1988). While class I membrane glycoproteins are expressed on almost all nucleated cells, class II molecules are expressed only on specialized antigen-presenting cells such as dendritic cells, activated B lymphocytes and macrophages (Muhlethaler-Mottet et al., 1997, Holling et al., 2004). The non-classical MHC class II molecules, DO and DM, are responsible for regulating the binding process of foreign peptides (Ting and Trowsdale, 2002).

The MHC harbours the most variable functional loci in vertebrates and there is a bulk of empirical evidence that positive selection acts on the MHC loci to maintain this variation (Piertney and Oliver, 2006). Polymorphisms in the MHC genes are associated with inflammatory, infectious and autoimmune diseases in humans as well as other mammalian species. In the horse, three immune-mediated diseases, Insect Bite Hypersensitivity (Schurink et al., 2012, Andersson et al., 2012, Klumplerova et al., 2013), Equine Sarcoids (Staiger et al., 2016) and Equine Recurrent Uveitis (Fritz et al., 2014) show significant association with genetic markers located in the MHC class II region.

The first complete sequence (3.6 Mb) and gene map of human MHC or human leukocyte antigen (HLA) was published in 1999 by the MHC sequencing consortium (The MHC sequencing consortium, 1999). Following the sequencing and annotation of human MHC, reliable MHC assemblies were generated for mouse and rat (Hurt et al., 2004), domestic dog (Debenham et al., 2005) and cat (Yuhki et al., 2007), cattle (Childers et al., 2006), pig (Ando and Chardon, 2006), panda (Wan et al., 2009) and sheep (Gao et al., 2010). Obtained sequences provided information for larger scale comparative evolutionary studies illustrating the high genomic plasticity of this region (Kelley et al., 2005, Wan et al., 2009, Takahashi et al., 2000).

Investigation of the horse MHC began with its assignment to the region on chromosome 20 q14-q22 (Ansari et al., 1988). In the following years studies describing mainly polymorphism in the second exon of the MHC class II genes were conducted and a considerable extent of polymorphism was observed (Albright-Fraser et al., 1996, Fraser and Bailey, 1998, Brown et al., 2004, Diaz et al., 2001, Gustafsson and Andersson, 1994). Horse *DRB* (Fraser and Bailey, 1996), *DQA* (Fraser and Bailey, 1998, Kamath and Getz, 2011) and *DQB* (Horin and Matiasovic, 2002) genes appeared to be encoded by multiple loci. According to RefSeq gene annotation of the horse reference genome, EquCab2 (Wade et al., 2009), two of the known *DRB* genes were not located in a tail-to-tail cluster with the *DRA* gene, as observed in human and mouse.

To complement the assembly of EquCab2, Children's Hospital Oakland Research Institute created a BAC library (Osoegawa et al., 1998) called CHORI-241 from a Thoroughbred breed stallion Bravo, a half sibling of the horse

Twilight that was used as a DNA source for the horse reference genome (Wade et al., 2009). In previous studies (Tallmadge et al., 2005, Gustafson et al., 2003) this BAC library was screened for MHC genes to assemble a minimal tiling paths of the whole MHC region and to obtain a relative gene map.

2.6 Evolution by gene duplication

Functional genomic loci are most commonly subjected to purifying selection that aims to remove deleterious mutations. As a result, unfit SNVs and small INDELs are subjected to purifying selection and contribute very little to genomic plasticity and phenotypic diversity of species. Gene duplication is a process that relaxes the constraint of the purifying selection (Ohno, 1970). As long as one gene copy remains functional, the other copy can accumulate genomic variation leading to subfunctionalization, neofunctionalization or loss of function.

There are two main mechanisms that can lead to gene duplication – retrotransposition and unequal crossing-over (Zhang, 2003). During retrotransposition mRNA from an expressed gene is fully retrotranscribed into a cDNA and further integrated into the genome, resulting in an intron-less copy of the gene or so called "retrogene" (Kaessmann et al., 2009). Unequal crossing-over can occur during meiosis due to the misalignment of regions with sequence similarity at the maternal and paternal chromosomes or sister chromatids of either maternal or paternal chromosomes (Lupski, 1998). As a result of unequal crossing-over, a segmental duplication containing exon-intron structure of the parental gene and, possibly, also regulatory features is produced (Wolfe and Li, 2003). Sequence similarity in distinct genomic regions can occur by random integration of TEs, repetitive microsatellite DNA or microhomology (Goldberg et al., 1983, Bailey et al., 2003, Barsh et al., 1983, Conrad et al., 2010). In addition, TEs can also cause unequal crossing-over directly, by alternative transposition (Gray, 2000). Alternative transposition is an alternative version of the traditional transposition of TEs that has been described in bacteria, maze, tobacco and drosophila (Gray, 2000) and can result in large scale structural rearrangements.

The most significant contribution to evolution of phenotypic diversity is caused by whole-genome duplication. Evidence for two rounds of whole-genome duplication in a common ancestor of the vertebrate genome has been described (Dehal and Boore, 2005). A third round of whole-genome duplication has occurred in teleost fish, the most species-rich group of vertebrates (Amores et al., 1998).

3 Aims of the thesis

The overall aim of this thesis was to explore the advances of massive parallel sequencing for studying genomic complexity in domestic animals and to establish pipeline for mapping disease-related traits.

The specific aims were to:

- Evaluate Ion Proton sequencing technology for whole-genome re-sequencing.
- Perform whole-genome sequencing of a family trio to investigate the genetic variants associated with an autosomal recessive disease in Kromfohrländer dogs denoted hereditary footpad hyperkeratosis.
- Resolve genomic structure of the horse major histocompatibility complex class II region and provide accurate annotation.
- Explore the evolution of the horse MHC class II haplotype.

4 Present Investigations

4.1 Study I

Evaluation of whole-genome sequencing of four Chinese crested dogs for variant detection using the Ion Proton system

The most common approaches for genomic variant detection is genotyping with SNP arrays or whole-genome re-sequencing with the Illumina platform. In this study we sequenced four Chinese crested dogs using the Ion Proton system to evaluate an alternative technology for whole-genome re-sequencing of canine genome for variant detection.

Coverage statistics

The average throughput of each PI chip was 9.5 Gb, corresponding to ~73 million single reads with the mean read length of 130 bp, even though the 200 bp fragment libraries were prepared. Fragment library from each dog was sequenced on two PI chips each. On average 98.5 % of the reads could be aligned to the CanFam3.1 reference sequence. Two types of coverage were analysed – "vertical coverage" describing average number of reads covering each base of the genome, and "horizontal coverage" describing a proportion of the genome covered with at least 4 reads. The average estimate of "vertical coverage" was 6x for all sequenced individuals. The average estimate of the "horizontal coverage" was 80 % of the whole genome and 77 % of the known exome. We detected a gradual drop in coverage if the GC content was less than 35 % and more than 60 %.

Each library preparation may introduce a certain random bias related to genomic content present in the particular library. To evaluate whether sequencing more libraries and chips would increase the “horizontal coverage” we performed a library merging analysis where coverage statistics were calculated after adding each extra sequencing library (data from 2 PI chips). The largest gain in coverage was the addition of the first extra library, where horizontal coverage increased up to 94.6 % genome wide and 90.8 % exome wide. The combination of all four libraries sequenced on 8 PI chips increased the “horizontal coverage” up to 97.2 % genome wide and 94.3 % exome wide, leaving 14.8 Mb of the genome or 1.3 Mb of the exome with zero coverage. These regions mostly corresponded to transposable elements, CpG Islands and regions with high GC content.

Variant detection and evaluation

With individual dog variant calling we detected on average 2.4 million SNV and 0.7 million INDEL positions per dog and with common analysis we obtained approximately 4 million SNVs, similar results for both variant detection tools, and 2.7 INDELs by UnifiedGenotyper (McKenna et al., 2010) and 5.6 million with SAMtools (Li et al., 2009).

Two of the sequenced dogs were also genotyped with the 170 K CanineHD BeadCip (Illumina). For both individuals, slightly more than 90 % of the SNV genotypes called by UnifiedGenotyper were concordant with the genotypes from the array. In 60 % of the miscalled genotypes re-sequencing analysis even with 5-fold average coverage have failed to identify the alternative allele that has been detected with the array.

Even though INDEL calling with Ion Proton data is challenging due to specific platform errors, in case of SNV calling it can serve as an alternative to other next-generation sequencing platforms and SNP genotyping arrays. In addition, we identified new genetic variants of the Chinese crested dog that will contribute as known canine genetic variation.

4.2 Study II

Whole-Genome Sequencing of a Canine Family Trio Reveals a *FAM83G* Variants Associated with Hereditary Footpad Hyperkeratosis

To map an autosomal recessive disease caused by a single gene mutation the traditionally used GWAS approach would require genotyping approximately 10-20 affected dogs and an equal number of healthy individuals (Karlsson and Lindblad-Toh, 2008). However, due to the large degree of linkage diequilibrium within domestic dog breeds, the associated regions are usually large, containing 0.5-1 Mb regions that need to be further re-sequenced.

Whole-genome sequencing (WGS) and whole exome-sequencing (WES) has been successfully used for identifying mutations associated with inherited Mendelian disease in human. Hereditary footpad hyperkeratosis (HFH) is a canine monogenic disease, presumably with a recessive inheritance pattern. To test disease mapping by trio sequencing in dogs, we chose a family trio from the Kromfohrländer breed, where the offspring was diagnosed with HFH and both parents were expected to be carriers of the disease allele. All three individuals were sequenced on the Ion Proton system, each individual on two PI chips.

Variant detection and filtering

The number of good quality variants from the whole-genome re-sequencing analysis of all three individuals was 3,449,902 SNVs and 198,165 INDELs. By taking advantage of the existing annotation of CanFam3.1 reference sequence (Hoeppner et al., 2014), we eliminated all variants present in noncoding regions and fractioned the exonic SNV variants into missense, nonsense and silent substitutions. Further, only the variants that followed the inheritance pattern (heterozygous in parents and homozygous in offspring) were retained. By comparing the remaining variants to the data base of known variants in dogs, (Axelsson et al., 2013) and Ensembl Variation Release 77 (Canis lupus familiaris), only 52 missense substitutions and no INDELs were left as candidates. Out of 52 candidates, only 17 variants were deleterious and after genotype validation we were left with seven candidate loci that were Sanger-sequenced in an additional eight HFH-affected and 16 healthy Kromfohrländer dogs. Only two loci, *GRAPL* and *FAM83G*, were homozygous for the variant allele in all cases and none of the control individuals.

The missense variant located in the *FAM83G* was a G to C transversion at position chr5:41,055,619 (CanFam3.1). This transversion resulted in a positively charged arginine to be exchanged with a neutral proline. In the 29 mammals' comparison (Lindblad-Toh et al., 2011) this arginine residue at p.52 was completely conserved. In contrast, two adjacent substitutions in *GRAPL* at position chr5: 41,014,230 - 41,014,231 were not conserved in the 29 mammals' data. This sequence of the genome was not annotated as protein coding in any other of the compared species, suggesting that it might be an erroneous annotation. In addition, an extra analysis of 48 additional unaffected dogs showed the presence of the homozygous variant genotype for both SNP positions in eight unaffected individuals.

We concluded that the missense variant FAM83G:c155G>C(p.R52P) is associated with HFH in Kromfohrländer dogs. A previously published GWAS and WGS of a single individual combination independently came to the same conclusion in two different dog populations (Drogemuller et al., 2014). Nevertheless, we have shown that family trio sequencing is a powerful technique for identification of rare alleles associated with Mendelian disorders, circumventing the challenge of gathering sufficient number of individuals needed to achieve statistical power in GWAS.

4.3 Study III

Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology

The repetitive nature of the Major Histocompatibility Complex (MHC) region, including segmental duplications, copy number variation and abundance of large transposable elements, is a major challenge for producing a reliable assembly over this region. A BAC library CHORI-241 was developed by researchers at the Children's Hospital Oakland Research Institute (CHORI) in California to supplement the whole genome shotgun sequencing of the horse reference genome EquCab2 (Wade et al., 2009). Using CHORI-241 BAC library a minimal tiling paths over the entire horse MHC region has been constructed and relative gene map has been previously described (Tallmadge et al., 2005, Gustafson et al., 2003).

The long-read SMRT sequencing technology from Pacific Biosciences has been successfully used to resolve complex genomic regions and to improve genome assemblies (Vij et al., 2016, Frank et al., 2016, Huddleston et al., 2014, Chaisson et al., 2015). Thus, to resolve the repetitive complexity and genomic structure of the horse MHC class II region, we sequenced eight BAC clones spanning MHC class II region with PacBio long-read sequencing technology and performed a comprehensive annotation of the assembled sequence.

Assembly and annotation

The final assembly of all BAC clones and two long-range PCR products resulted in 1,165,328 bp continuous gap-free sequence spanning the classical MHC class II region and part of the upstream "gene desert" (Figure 2). An automated annotation pipeline analysis and further manual curation of the predicted coding loci resulted in 35 identified genes in the final horse MHC class II assembly. Of the detected loci, 23 had undisrupted reading frames and 12 were pseudogenes.

We identified coding genes for all types of classical MHC class II antigen presenting molecules. However, only two non-functional DP gene remnants were found, *Eqca-DPB1* and *–DPB2* pseudogenes. As expected, we identified three functional *Eqca-DRB*, three *Eqca-DQA* and three *Eqca-DQB* loci that have been described before. In addition to functional genes, we also located three *Eqca-DRB*, a single *Eqca-DQA* and a single *Eqca-DQB* pseudogene. We also identified all expected non-classical MHC-class II genes, such as *Eqca-DOA*, -*DOB*,-*DMA*, and -*DMB*). Surprisingly, two additional Eqca-DOB pseudogenes were found. During the annotation process, few horse-specific splice site mutations were detected (*Eqca-DRB3*, *-DOA*) and, a particularly interesting case, when after the 3'end and final exon deletion of *Eqca-DQB3* gene*,* the transposable element inserted downstream of the deleted sequence was used to produce an alternative ending of the *Eqca-DQB3* transcript.

Comparison to the EquCab2

The comparison of our final MHC class II assembly and current EquCab2 reference genome did not reveal any major structural discrepancies, but pinpointed one gap and one single inverted duplication of a sequence corresponding to the transposable element. Nevertheless, the sequence comparison revealed that half of the Bravo MHC class II region (up to *Eqca-DRB2* gene) was very similar to Twilight's haplotype, while the number of SNVs and INDELs escalated drastically starting with middle point of the *Eqca-DRB2* gene. We imply, that even though seldom INDEL errors can persist, the deep coverage of PacBio sequencing, the majority of observed discrepancies

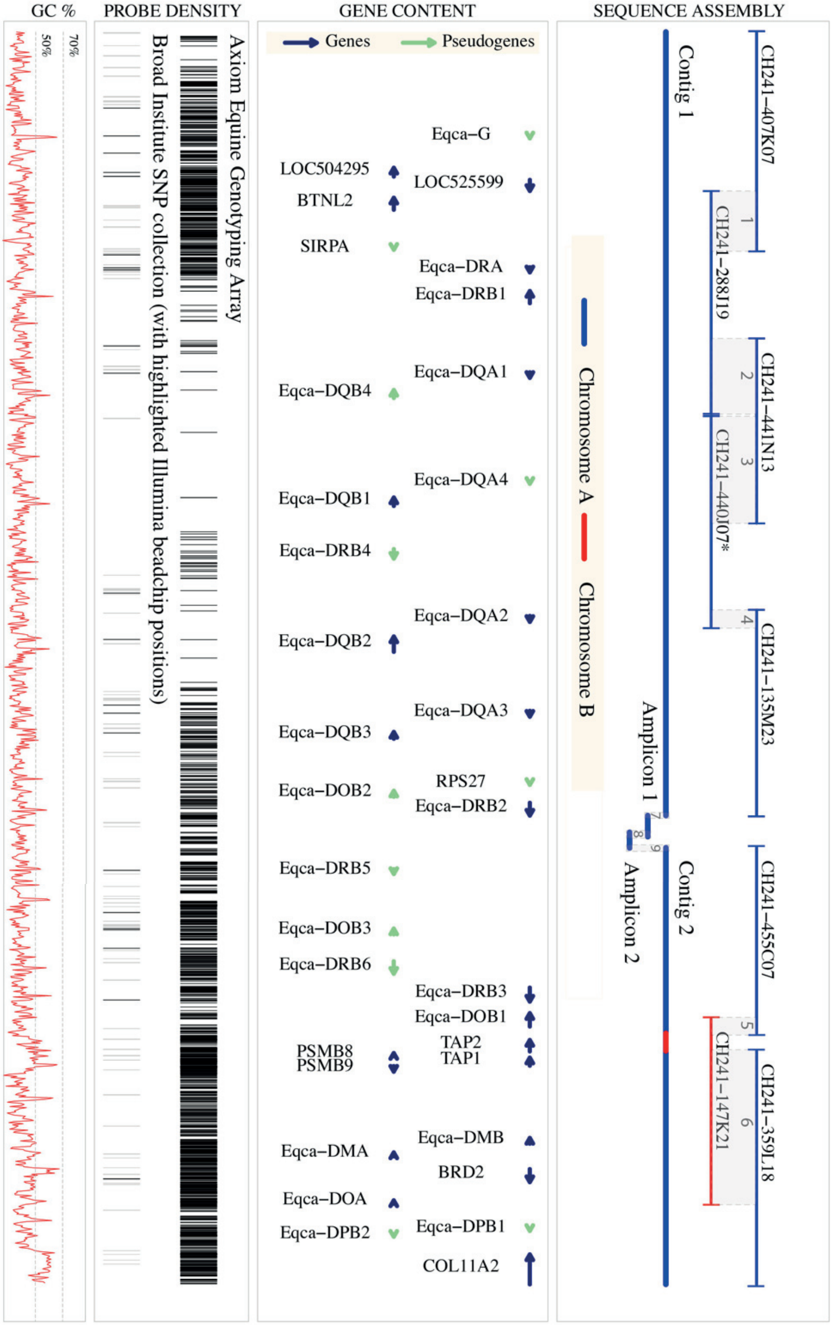

Figure 2. Physical map of the horse MHC class II region (Viļuma et al., 2017)

between these two sequences represented a combination of EquCab2 sequencing errors and *bona fide* sequence variation.

MHC class II structure in mammals

By comparing horse MHC class II structure to six other mammalian MHC class II regions (Figure 3), we identified both, regions of conserved gene order and directionality and regions of considerable structural variation. Most of the structural variation was shown to exist in the DR and DQ gene families. Peculiarly, in contrast to the organization of *DRB* genes in other compared mammals, six *Eqca-DRB* loci were distributed across a region spanning ~0.7 Mb with intervening *DOB* and *DQ* loci. Interestingly, all of the compared Laurasiatheria mammals had a single inverted *DRB* gene, while this was not observed in human and mouse. Even though the phylogenetic analysis of protein-coding genes suggested species-specific evolution of the paralogous copies, it is intriguing to speculate that this inversion might have occurred 85-100 million years ago, after the separation of Euarchontoglires clade (represented by human and mouse) but prior to the split of Laurasiatheria (represented by five domestic animals) into major orders.

In conclusion, PacBio long-read sequencing was sufficient for providing the first in-depth sequenced gap free horse MHC class II sequence. A continuous, well-annotated "reference" sequence is essential for better experimental design of equine immune-mediated association studies, like Insect Bite Hypersensitivity, Retinal Uveitis, Equine Sarcoids and possibly many other allergic diseases. Horse MHC class II is also the first high quality assembly and annotation of an odd-toed ungulate order and therefore provides a valuable piece of the puzzle for understanding the genomic plasticity of the MHC class II region in domestic species.

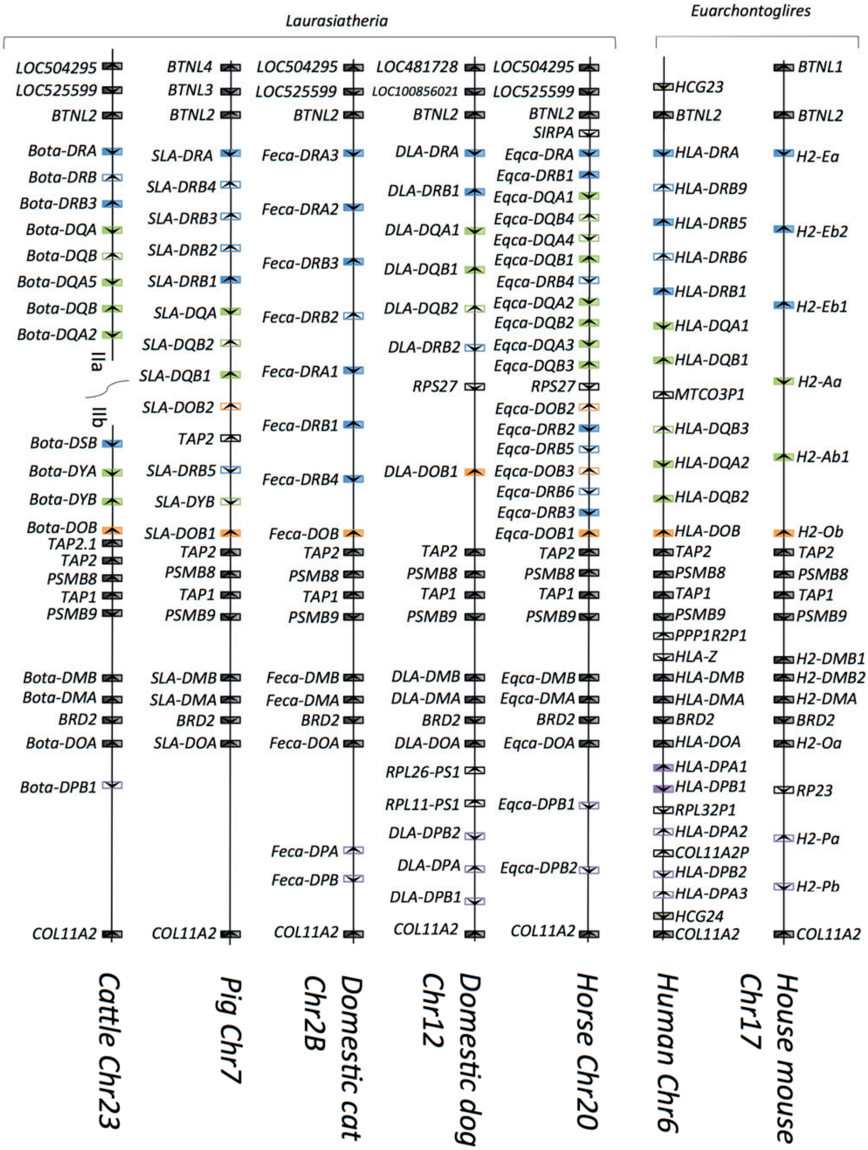

Figure 3. The schematic gene order comparison of the horse, mouse, human, cat, dog, pig and cattle MHC class II structure (Viļuma et al., 2017). The conserved genes are coloured in black and other loci are coloured according to gene family they belong to. Filled boxes represent functional genes and empty boxes represent pseudogenes. For better overview, the cattle MHC class IIb sequence was inverted.

4.4 Study IV

Evolution of the horse MHC class II region by gene duplication

In Study III we showed that, in comparison to six other mammals (human, mouse, pig, cattle, domestic cat and dog), the horse MHC class II region distinguished itself with the most extraordinary structure in terms of increased number of paralogous copies, including five inverted *DRB* genes, and by the lack of gene clustering according to the type of MHC class II molecule, as *Eqca-DRB* genes were found in four different locations separated by Eqca-*DQA*, *-DQB* and *-DOB* loci.

The extraordinary structure together with the high depth long-read assembly provided an excellent opportunity for studying evolution by gene duplication and the putative role of large transposable elements in the molecular evolution and plasticity of horse MHC class II region. In this study we investigated the size and the age of the duplicated segments, as well as TEs residing at the homology break points to facilitate a better understanding of the evolutionary mechanisms that have shaped horse MHC class II region.

The structure of segmental duplications

We analysed the genomic region from the 3' end of *Eqca-DRA* to the 5' end of the *TAP2* gene in the horse MHC class II region, that included all paralogous *DRB*, *DQA*, *DQB* and *DOB* genes. The "self-comparison" analysis confirmed the highly repetitive structure of this region, where the majority of genomic sequence has originated from several tandem duplications of different size and non-tandem segmental duplications, leaving only 1.6 % of the sequence as a single copy. The largest duplicated segments were four segments containing tail-to-tail located *Eqca-DQA* and *-DQB* genes.

The conservation of exon-intron structure in the duplicated genes suggests that duplication most likely occurred through DNA-mediated events like unequal crossing-over rather than through reverse transcription (Wolfe and Li, 2003). Exon-intron structure was observed in all DQ, DRB and DOB segments suggesting that the gene duplication of these genes occurred through unequal crossing-over.

Phylogenetic relationship of paralogous segmental duplications

Multiple sequence alignments and maximum likelihood (ML) clustering of six DRB segments suggested that the initial expansion was an ancient inverted duplication event of the *DRB1* gene approximately 90 million years ago (MYA). The consecutive expansion of the inverted DRB segment was dated to approximately 55-70 MYA and resulted in presence of five inverted DRB segments. The similar analysis of DQ and DOB segments showed that these duplication events coincided with the expansion of inverted DRB segments. This time period overlaps with the mammalian radiation and further expansion of Laurasiatheria orders.

Analysis of transposable elements

In total, TEs constituted 61 % of this genomic region and the most common TE types found were equine repetitive elements (ERE) from the SINE family, and L1 elements from the LINE family. The homology breakpoints of nearly all segmental duplications were dominated by large LINE sequences, in particular L1MAB_EC, different L1-5 types, L1ME and L1MC elements.

In this study we have investigated the repetitive structure of the horse MHC class II region and described segmental duplication events of different size and nature that have shaped the known haplotype in horse. Even though the major expansion of this region was estimated to coincide with mammalian radiation, the initial duplication of the DRB segment might have occurred before the Laurasiatheria split in main orders. We showed that the high genomic structural diversity and plasticity of the horse MHC class II region may be a result of recurring segmental duplication events mediated either by unequal crossing-over between misaligned genomic regions or, possibly, by alternative transposition of long interspersed nuclear elements.

5 General Discussion

The increase in number of sequenced eukaryotic genomes clearly illustrates the level of contribution of massive parallel sequencing in the characterization of the genomic complexity. Without a doubt, human genome sequencing is the leading provider of the whole-genome data. In 2015 the 1000 genome project consortium published a global reference for human genetic variation, obtained by sequencing of 2504 individuals (The Genomes Project, 2015). While a year later, the report of deep sequencing of 10 000 human genomes identifying over 150 million variants was published (Telenti et al., 2016). In a decade of high throughput sequencing, a variety of methods has been adopted to study not only genomes, but also eukaryotic transcriptomes (Wang et al., 2009), gene regulation (Farnham, 2009) and epigenetic mechanisms (Park, 2009, Laird, 2010). The research questions have covered a broad variety of topics from curiosity-driven population history investigations to disease mapping, diagnosis and precision medicine.

High throughput sequencing has also significantly contributed to genomic exploration in domestic species. The number of re-sequenced domestic animal genomes can already be measured in the thousands as large scale genomic projects, such as the 1000 bull genomes (http://www.1000bullgenomes.com), dog 10K genome project (http://www.dog10kgenomes.org), 99 Lives cat genome sequencing initiative (http://felinegenetics.missouri.edu/99lives) and recently a Nordic dog breed sequencing project has been launched and several individual research groups have taken great advantage of massive parallel sequencing for phenotypic trait mapping.

Initially, whole genome re-sequencing was used to further follow up genome-wide association results, to pinpoint the causing mutation in the large associated regions (Frischknecht et al., 2013, Owczarek-Lipska et al., 2013, Jagannathan et al., 2013). Studies I and II contributed knowledge on the use of the Ion Proton platform for dog genome re-sequencing analysis and showed that the whole-genome trio-sequencing can lead to the discovery of causative mutations without prior GWAS in dogs. The genotype validation with the SNP array suggested the necessity of the higher sequencing depth in order to increase detection of the alternative allele of the heterozygous single SNV positions. However, the higher read coverage would compromise more individual genomes being sequenced, as the resources are limited. According to low-coverage sequencing statistics from Study I, the obtained data throughput would be sufficient to cover approximately 77 % of the exome with four or more reads and called SNV genotypes would have 10 % chance of wrong genotype. Nevertheless, in Study II we performed low-coverage sequencing of three individuals as an initial genome scan for non-synonymous mutation segregating with the expected inheritance pattern. Despite the high probability of missing causative variant due to low-coverage, we were able to pinpoint a single mutation in the *FAM83G* locus. However, in this particular case, the fact, that an independent research group reached the same conclusions (Drogemuller et al., 2014), provided reliability for the results. In absence of such information, a question could be raised, whether the additional variation with the same inheritance pattern could have been lost due to poor coverage in 23 % of the exome. In such circumstances, extra sequencing would be preferable to ensure the robustness of obtained results.

Recently, it was shown that the sequencing of a single individual was sufficient to find a genomic variant responsible for a phenotype when prior knowledge of candidate genes existed (Caduff et al., 2017, Durig et al., 2017). Considering the constantly accumulating information in other vertebrate species, it is important to highlight the important role of comparative analysis to study vertebrate biology, evolution and to transfer the existing knowledge between the species (Meadows and Lindblad-Toh, 2017).

The impact of long-read sequencing technologies

Development of single molecule long-read sequencing technologies, such as PacBio and ONT, have made it possible to successfully resolve complex, highly repetitive genomic regions and phase the variation from diploid organisms (Huddleston et al., 2014, Huddleston et al., 2017, Roe et al., 2017, Hans et al., 2017). High quality *de novo* assemblies can be generated primarily using PacBio (Gordon et al., 2016, Jiao et al., 2017, Vij et al., 2016) or ONT (Jain et al., 2017) sequencing technologies. With long-read sequencing it has become possible to investigate the parts of genome that have been hidden from short-read technologies. This includes detecting and understanding mechanics of large scale, complex structural changes and copy number variation. A good example illustrating the superiority of this technology was highlighted in a study where long-read sequencing of human haploid genomes revealed approximately 89% more variants than it has been reported in 1000 Genomes Project (Huddleston et al., 2017). Furthermore, long-read sequencing is gaining a deserved appreciation in precision medicine (Ashley, 2016, Huddleston et al., 2017) and, recently, the first success story of long-read sequencing in diagnostics of a Mendelian disease has been published (Merker et al., 2017).

Studies III and IV are supportive examples of the benefits provided by long-read sequencing. The MHC class II region harbours some of the most polymorphic multi-gene families with copy number variation. The repetitiveness of this region, including segmental gene duplications and high density of long interspersed elements, was an excellent challenge to be solved with the long-read technology. In Study III we performed PacBio sequencing of eight BACs and produced a continuous 1.2 Mb assembly spanning the horse MHC class II region. This gap free assembly served as a robust backbone for gene annotation and allowed identification of 35 genomic loci, including gene families encoding for the MHC class II antigen presenting molecules. Results confirmed the unusual directionality of the *Eqca-DRB2* and *-DRB3* loci and revealed an additional peculiarity - increased number and scattered *Eqca-DQ*, *-DRB* and -*DOB* loci in horse. The reliability of the long-read assembly allowed the inference of the evolutionary history of this region by analysing segmental duplication events and transposable elements found at the homology break points.

Importance of high-quality annotation

The human and mouse genome annotation are good examples of a comprehensive set of information that has been collected to facilitate the understanding of the genomic complexity. In human, most of the re-sequencing studies have been aimed exclusively at the exome, reducing the cost per sample and increasing coverage depth. The key factor for the successful outcome of the targeted exome sequencing is a reliable capture chemistry, developed based on a comprehensive annotation. In domestic animals there are few studies (Ahonen et al., 2013, Cosart et al., 2011, McClure et al., 2014) that have used exome capture approach. The most comprehensive available exome capture assay in domestic species is based on the latest reference genome assembly of the domestic dog, CanFam3.1 (Broeckx et al., 2014). This assay captures 2 % of the genome, based on information in the Ensembl Genes, the RefSeq Genes, as well as mRNA and non-coding RNA annotation (Broeckx et al., 2015) and it has been successfully used to determine the causative variants in several dog breeds (Cox et al., 2017, Tsuboi et al., 2017, Evans et al., 2016). Overall, the annotation in domestic species requires considerable improvements and additional layers of complexity, such as promoters, enhancers, transcription factor binding sites and non-coding RNAs. To address this issue, a large collaborative initiative The Functional Annotation of Animal Genomes (FAANG) has been established (http://www.faang-europe.org). The aim of the FAANG project is to deliver standardized data sets from individual animal species and to establish an infrastructure for analysing genome-wide functional data.

Sequencing of Major Histocompatibility Complex

Study III was a valuable experience in understanding the importance of the fit between technology and the research question. For the initial attempt of MHC sequencing, we chose two horses, a Swedish Warmblood and the Faroese Pony, that were homozygous for microsatellite markers over the MHC class II region. Both libraries were sequenced with Ion Proton platform and reads of average size of 120-130 bp were aligned to the EquCab2 reference genome. In contrast to the expected homozygosity, both individuals showed heterozygous locations and, in Faroese Pony, the number of alleles at *Eqca-DRB1* and *-DQA* loci varied from zero to three alleles (Figure 4A). The issue could not be entirely solved by Illumina paired-end read, nor SOLiD mate-pair (spanning 10 kb inserts) sequencing, as the length of mate-pair reads were only 75 bp in lenght (Figure 4B). As a consecutive next step, long-read approach was chosen for sequencing the BAC clone minimal tiling paths over the MHC class II region. Both long-read sequencing technologies, PacBio and ONT, were tested. However, analysis was performed when ONT was at its initial stage of development with insufficient throughput and high error rate. Finally, PacBio sequencing was sufficient for the reliable assembly of the MHC class II region.

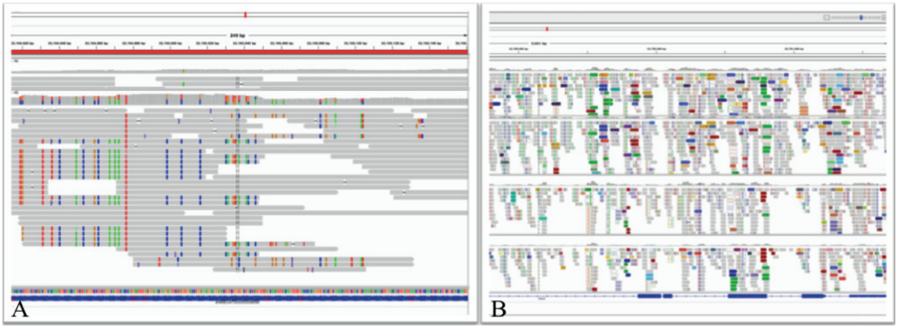

Figure 4 Inconclusive read mapping of A) Ion Proton sequence data from whole-genome re-sequencing to *Eqca-DQA3* exon 2 (3 alleles present) and B) uneven coverage and faulty mapping of SOLiD 10 kb mate-pair reads to the *Eqca-DRB1* gene.

6 Concluding remarks and future perspectives

The ultimate goal of research of domestic animals is to improve their health and welfare, while increasing productivity or performing selective breeding of desirable exterior and behaviour traits. Understanding the link between phenotypes and genomic variation is an essential step towards this goal. Therefore, “a perfect genomic study” is the one that can identify the causative mutation with minimal time and financial costs and by using as few samples as possible.

The proof-of-concept study of trio sequencing (Study II) together with numerous other success stories of mapping causative disease variants in horse, cat, dog etc. is good proof that whole-genome/whole-exome sequencing is a powerful tool for causative variant discovery in monogenic diseases and traits in domestic animals. It also confirms that the methodology for this approach is already well adopted in domestic animal research and suggests that in coming years a “Great Flood” of studies pinpointing genomic variation behind single-traits will continue. Certainly, long-read sequencing will play an important role in revealing the effects of structural variation on a phenotype.

It has been shown that sequencing of a single individual is sufficient to detecting the causative variants in both, humans and domestic species. However, to increase the success rate of such studies, there are several important issues that need to be addressed in order strengthen the pillars of success. First of all, “finished” high quality reference genomes are fundamental. Improved reference sequence assemblies taking advantage of long-read sequencing will soon be available for domestic cat and horse, and, hopefully, the others will follow shortly. Comprehensive annotation is another important aspect. While there are several initiatives to add extra layers of annotation, such as non-coding RNAs, transcription factor binding sites and others, it is important to note that protein-coding gene annotation in domestic species is far from perfect and requires improvements. Data sharing, standardized protocols, catalogues of known variation, and large scale collaborative initiatives, such as FAANG and the 200 mammal sequencing project, are expected to result in substantial improvements of genome annotation. Another important aspect lays in comparative genomics and knowledge that is already present in other species. It is essential to catalogue all identified mutations, phenotypic traits and diseases in a way that is easily accessible by automated pipelines. Last, but not least, the correct assessment of the phenotypic traits or disease diagnosis is essential and collaboration with veterinarians, epidemiologists and breeders should not be underestimated.

Regardless of direct economical and social value, domestic animals can provide general knowledge in evolution and biology, and serve as model organisms for understanding human traits. Each well-characterized genome of a domestic animal is additional information source in our understanding of the general picture of mammalian genomic complexity and its evolution.

Long-read sequencing technologies have made a great promise to science. However, the current throughput of long-read technologies is not optimal and sequencing of the entire genome remains costly. The recent launching of new Sequel instrument by Pacific Biosciences is expected to increase the throughput by a magnitude of seven. Nevertheless, it would be beneficial to direct long-read sequencing for studying only the parts of the genome that requires long-read analysis, *i.e.*, complex regions and repetitive fractions of DNA. It has been shown before that it is possible to eliminate the repetitive fraction from eukaryotic genomes to solely capture the unique sequences for library preparation (Peterson et al., 2002). An inversed version of this approach could be used to aim long-read sequencing towards moderately and highly repetitive fractions of the genome. Targeted capture of large genomic regions is another prerequisite for efficient long-read studies. Few capture methods have been proposed, such as CATCH-seq (Day et al., 2014), enChIP using CRISPR (Fujita and Fujii, 2013) and Cas9-assisted targeting of chromosome segments (Gabrieli et al., 2017) .

It is exciting to see the amount of novel information that is hidden in the repetitiveness of the DNA and how much of additional phenotypic variation will be explained with short-read and long-read sequencing technologies. It is expected that functionally important elements such as centromeres and telomeres will be explored in depth and new insights in aging and longevity could be revealed. In addition, detailed investigations on structural variation of regions showing high genomic plasticity will aid in a deeper understanding of mammalian evolution.

Popular science summary

Life on Earth is characterized by an outstanding phenotypic diversity resulting from adaptations to a vast variety of environmental conditions that have ensured survival of living organisms for billions of years. Exploring genetic factors behind this variation has been a passion and long-term commitment of many scientists. After DNA was discovered as the true "blueprint" of life, a large number of studies have been conducted to understand the structure and information of the DNA molecule. In particular, the ability to determine the sequence of the building blocks of the four nucleic acids, also called nucleotides, has been instrumental for making phenotype-genotype correlations. The average mammalian genome consists of two to three billion of these nucleotides and only in the last ten years the sequencing technologies have allowed cost-effective large-scale genome sequencing projects. The traditional sequencing technology, called Sanger-sequencing, is capable of reading the order of only 1000 nucleotides in a single experiment. Therefore, 2 to 3 million Sanger-sequencing experiments would be necessary for reading the entire mammalian genome. In contrast, massive parallel sequencing technologies are capable of analysing up to a billion fragments of DNA molecules simultaneously (in parallel). Thus, in a single sequencing experiment it is possible to obtain the nucleotide order of the entire genome in a form of fragmented pieces, the size of which may vary from 75 to tens of thousands of nucleotides depending of the chosen technology. This thesis describes four studies implementing massive parallel sequencing technology to identify genetic variants associated with the development of disease and to characterise complex genomic region important for disease resistance in horse.

In the first study we evaluated IonProton sequencing platform by re-sequencing the entire genome of four Chinese Crested dogs. We compared all the generated nucleotide sequences to the already known, good quality, 2.4 billion bases long genome sequence of the dog, called the reference genome. The comparison revealed that out of 2.4 billion bases analysed, five million single bases were different and six million bases were either inserted extra or missing in at least one of analysed dogs. The second study, the whole-genome re-sequencing of a canine family trio from the Kromfohrländer breed, served as a proof-of-concept for a new approach to identify genetic causes of monogenic disease in dogs. In this study, we were able to narrow down the observed genomic variation of 3.5 million bases to a single nucleotide variant in *FAM83G* as a most likely cause of the autosomal recessive disorder termed hereditary footpad hyperkeratosis.

The other two studies highlighted the advantages of the technology that is capable of generating considerably longer nucleotide fragments, a method called single-molecule real-time (SMRT) sequencing. SMRT sequencing was chosen to resolve the genomic sequence of the horse Major Histocompatibility Complex (MHC) class II region. The MHC is a highly repetitive, gene dense genomic region that encodes proteins with an important role in immune defence against pathogens and is associated with resistance and susceptibility of various infectious, inflammatory and autoimmune diseases. Using SMRT sequencing we successfully constructed a more than one million base long continuous sequence spanning the horse MHC class II region. Further analysis of this sequence revealed precise location of 35 genes, including a high number of genes encoding the MHC class II antigen presenting molecules. The comparison of the horse MHC class II region to those of human, mouse, cattle, pig, domestic cat and dog, contributed a supporting knowledge of the incredible genomic plasticity of this region and provided a reliable foundation for further evolutionary and disease mapping studies.

The fourth study of the thesis showed how the high quality DNA sequence of the horse's MHC class II region, facilitated studies of the evolutionary history of this complex genomic region. The long-read sequence assembly provided an in-depth detailed picture, in which the order of genes and segments of sequences have been duplicated and inverted to form the complex, repetitive nature of MHC class II region in horse. We estimated that the majority of duplications of genes and segments coincided with the time of mammalian radiation, about 55-70 million years ago, with an exception of the ancient initial duplication-inversion of the Eqca-*DRB1* gene, that possibly existed already in the common ancestor of all domestic animals. In addition, we propose a hypothetical model where long repetitive elements, also known as "jumping genes", are factors behind the genomic plasticity of this region.

Populärvetenskaplig sammanfattning

Livet på jorden har anpassat sig till en varierande miljö vilket medfört en enastående fenotypisk mångfald som säkerställt överlevnaden av organismer under miljarder år. Många forskare studerar med passion och långsiktigt engagemang de genetiska faktorer som orsakar denna variation. Sedan upptäckten att deoxyribonukleinsyran (DNA) är den molekyl som utgör livets "textmall" (genomet), har en stor mängd studier genomförts för att förstå dess struktur och informationsinnehåll. ”Bokstäverna” i DNAt består av de fyra nukleotiderna adenin (A), cytosin (C), guanin (G) och tymidin (T). Ett däggdjursgenom består av två till tre miljarder nukleotider och den tekniska utvecklingen för att bestämma ordningen och variationen av nukleotiderna, har varit avgörande för att korrelera fenotyper med genotyper. Med de första sekvenseringsteknikerna kunde forskarna endast läsa ordningen av knappt 1000 nukleotider åt gången och därför skulle flera miljoner sekvenseringsexperiment vara nödvändiga för att läsa ett helt däggdjursgenom. De senaste tio åren har sekvenseringsteknologin utvecklats så att varje nukleotid i genomet kan analyseras med s.k. massiv parallell sekvensering där flera miljarder fragment av DNA-molekyler läses av samtidigt. I ett enda sekvenseringsexperiment är det därför nu möjligt att analysera sekvensen av ett helt genom. Denna avhandling beskriver fyra studier där massiv parallell sekvenseringsteknik har använts för att identifiera genetiska varianter associerade med sjukdomsutveckling, samt för att karakterisera en komplex genomisk region, viktig för sjukdomsresistens respektive -mottaglighet hos häst.

I den första studien utvärderade vi en metod med sekvenseringsinstrumentet IonProton för att att sekvensera hela genomet av fyra kinesiska nakenhundar. Vi jämförde DNA-sekvenserna från de fyra hundarna med hundens officiella referensgenom. Jämförelsen visade att av de 2,4 miljarder analyserade nukleotiderna, fanns skillnader i fem miljoner (2%) av dessa, medan sex miljoner (2,5%) var antingen ”insertioner” (extra nukleotider) eller ”deletioner” (nukleotider som saknas) i minst en av de fyra analyserade hundarna. I den andra studien användes samma sekvenseringsteknik för att finna orsaken till en ärftlig sjukdom hos hundrasen kromfohrländer. Hundar från denna hundras drabbas i högre grad än andra av en autosomalt recessiv sjukdom som kallas digital hyperkeratos (förhårdnader och sprickor i trampdynorna). För att identifiera orsaken till sjukdomen, helgenom-sekvenserades en familj bestående av en far, en mor och deras sjuka valp. Bland de 3,5 miljoner variabla nukleotiderna hos kromfohrländerfamiljen, kunde vi begränsa den sannolika genetiska orsaken till en enda nukleotidvariant i genen *FAM83G*.

Den tredje studien i avhandlingen visar fördelarna med en ny sekvenseringsteknik som kan generera betydligt längre nukleotidfragment än tidigare. Metoden kallas singelmolekylär realtidssekvensering (SMRT). De långa nukleotidfragmenten ger mycket stora möjligheter att läsa DNA-sekvensen i komplexa delar av genomet där repetitiva sekvenser och duplicerade gener gör det svårt att läsa sekvensen. Ett exempel på en sådan region är den som kodar för transplantationsantigener, ”Major Histocompatibility Complex” (MHC) klass II som har en viktig roll i det adaptiva immunförsvaret och är förenat med resistens eller mottaglighet för infektiösa, inflammatoriska och autoimmuna sjukdomar. Med SMRT-sekvensering har vi för första gången framgångsrikt konstruerat en lång (1,2 miljon nukleotider) kontinuerlig DNA-sekvens över hästens hela MHC-klass II-region. Den bioinformatiska analysen visade den exakta lokaliseringen av 35 olika gener, av vilka många är funktionella och kodar för MHC-klass II antigenpresenterande molekyler på cellytan. En jämförelse mellan hästens MHC klass II-region och motsvarande regioner hos människa, mus, ko, gris, katt och hund bidrog till ökad kunskap om den höga genomiska plasticiteten i regionen och gav dessutom en tillförlitlig grund för fortsatta evolutionära studier, och för fortsatt kartläggning av sjukdomar associerade med genetisk variation i MHC klass II.

Den fjärde studien i avhandlingen visar hur den nu producerade högkvalitativa DNA-sekvensen av hästens MHC klass II, underlättar studier av denna komplexa genomregions evolutionära historia. De långa sekvenserade fragmenten ger en tydligare bild än vad som tidigare varit möjlig, och i vilken ordning gener, och segment av sekvenser har duplicerats, inverterats o.s.v., i hästens MHC klass II. I studien föreslår vi en modell för hur dupliceringar av gener och segment skett i samband med tiden för däggdjurens evolutionära expansion för 55-70 miljoner år sedan. Hästens MHC-gen *DRB1*, var den första att dupliceras för mer än 90 miljoner år sedan. Dessutom föreslår vi en hypotetisk roll för hur långa repetitiva element, s.k. "hoppande gener", påverkar den genomiska plasticiteten i denna region

Acknowledgements

I would like to start my floods of thanksgiving by acknowledging **my three super wise supervisors**. You are a great team together and it has been an adventure to be part of it!
I will always be grateful to all three of you for introducing me to the MHC madness. I will never be able to see the world without the "MHC glasses" again.

Tomas, I won't write long and will go straight to the (bullet)point!
My biggest thanks to you for:

- ➢ My scientific development
- ➢ Evolution (it is difficult, but I absolutely love it)
- ➢ Scientific freedom
- ➢ Critical thinking
- ➢ Encouragement
- ➢ Filling in the "the"
- ➢ And all the wine and champagne!

Sofia, I would like to add, that it has been a great pleasure to work with you. Your endless optimism and excitement about science have encouraged me through all these years. You have an amazing gift to notice good things in every situation.

Göran, I hope, some day, I will now as much as you know. My big thanks to you for sharing your broad knowledge with me, starting from transposable elements and ending with bird songs. If one could combine all the time in the world, it still would not be enough to tell all your stories.

Thank you, **Marcel**, for being my bioinformatics advisor. Your advice and support has been truly appreciated and the technical tricks I have learned from you have made my life so much easier.

I would like to acknowledge **SciLifeLab**, **Uppmax** team and the **Bioinformatics Advisory program** for providing all the necessary infrastructure, support, education and scientifically creative environment. Also, a big thanks to Björn Nystedt and the rest of the inspiring Bioinformatics Advisory team of mentors and students who made me feel like I am a part of a creative, innovative and daring community.

Jag vill framföra min stora tacksamhet till **Stiftelsen Hästforskning** för finansiellt stöd av forskningen av hästens MHC till **Agria och Svenska kennelklubben** som varit med att finansierat hundforskningen och projektet om digital hyperkeratos.

I have also appreciated all the help from nice HGEN-Lab staff, **Louise**, **Charlotte**, **Susanne**, **Alexander** and **Siw**. Thank you, **Daniella**, not only for endless sequencing reactions, BAC clone extractions and the super long PCRs, but also for your friendship and all the wild adventures we had!

I would also like to kindly thank each and every one of my **colleagues at SLU** for being helpful, supportive, challenging, playful, encouraging, sometimes provocative (yes, you know who you are!), but most importantly, fun! **Cano**, tack för alla värdefulla svenska fraser. Ångrar ingenting som jag har gjort!

Especially, I would like to acknowledge **Susanne Eriksson** for being an amazing director of postgraduate studies and for your support when I needed it the most. **Fernando**, all the best of luck herding those cats! And by the way, thank you for Danny Bhoy!

Dear **past, present and future PhD students**, it has been my pleasure to know you, I have learned a lot from you and I will never be able to forget any of you! I should warmly emphasise two wonderful girls, who have tolerated me as an office mate for such a long time. **Merina** and **Suvi**, I have enjoyed sharing my scientific (and not so scientific) space with you!

Thank you **Nancy**, **André**, **Berihu**, **Ahmed**, **Thu**, **Sandrine, Bingjie and Josh** for all the horribly wonderful singing we have done!

And **Kim**: Brynäs! Brynäs! Brynääääs!

Hadrian, officially, thanks for taking care of planetsmasher and also for all the "black slope and lot of trees" fun!

Chrissy, thank you so much for tasty carrot salad last night! And banana muffins the night before! And vegan cookies the night before that! … I would be skin and bones without you! Thank you for taking care of me!

Also, I would like to send warm greetings to all **my friends at BMC** that did not gave up on me just because I moved to study to “countryside”: my sweet **Iris**, **Jonas**, **Freyja** (thank you for writing really nice thesis!), **Fabiana**, **Nima**, **Sangeet**, **Axel**, **Max**, **Fan**, **Tabea**, **Ravi** and **to many others**.
Jonas, thank you for letting me bother you with recombination and alignment questions even while being on holidays, and thank you for the PRANK!
Marta and **Reto**, you are like my MHC godparents. Hope to meet you some day again!

There are so many important people who have filled my PhD years with sports, adventures and countless parties. Among them my **volleyball friends**, **Gunnar**, **“Meetup” gang** and, of course, **UARS**!!! My dear Leksand rowing team, **Erik**, **Lutz** and **Stephan**, thank you for preventing me from working late nights and weekends, by luring me into midnight rowing, fun training and all the challenging Leksand races. I have missed you all this last year!

Liels paldies arī visiem maniem **Upsaliešiem** par to visu latvisko, ko mēs kopā esam piedzīvojuši!

Kā arī milzīgs paldies **Zibenītim** par dejām, jautrību un modra gara uzturēšanu!

Visbeidzot, paldies **mammai** un **tētim**, tā vienkārši – PAR VISU!

References

AHONEN, S. J., ARUMILLI, M. & LOHI, H. 2013. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. *PLoS ONE,* 8**,** e72122.

ALBRIGHT-FRASER, D., REID, R., GERBER, V. & BAILEY, E. 1996. Polymorphism of DRA among equids. *Immunogenetics,* 43**,** 315-317.

AMORES, A., FORCE, A., YAN, Y. L., JOLY, L., AMEMIYA, C., FRITZ, A., HO, R. K., LANGELAND, J., PRINCE, V., WANG, Y. L., WESTERFIELD, M., EKKER, M. & POSTLETHWAIT, J. H. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science,* 282**,** 1711-4.

ANDERSSON, L. S., SWINBURNE, J. E., MEADOWS, J. R., BROSTROM, H., ERIKSSON, S., FIKSE, W. F., FREY, R., SUNDQUIST, M., TSENG, C. T., MIKKO, S. & LINDGREN, G. 2012. The same ELA class II risk factors confer equine insect bite hypersensitivity in two distinct populations. *Immunogenetics,* 64**,** 201-8.

ANDO, A. & CHARDON, P. 2006. Gene organization and polymorphism of the swine major histocompatibility complex. *Animal Science Journal,* 77**,** 127-137.

ANSARI, H. A., HEDIGER, R., FRIES, R. & STRANZINGER, G. 1988. Chromosomal localization of the major histocompatibility complex of the horse (ELA) by in situ hybridization. *Immunogenetics,* 28**,** 362-4.

ASHLEY, E. A. 2016. Towards precision medicine. *Nat Rev Genet,* 17**,** 507-522.

AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. 1944. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. *J. Exp. Med.,* 79**,** 137-158.

AXELSSON, E., RATNAKUMAR, A., ARENDT, M. L., MAQBOOL, K., WEBSTER, M. T., PERLOSKI, M., LIBERG, O., ARNEMO, J. M., HEDHAMMAR, A. & LINDBLAD-TOH, K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature,* 495**,** 360-4.

BAILEY, J. A., LIU, G. & EICHLER, E. E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet,* 73**,** 823-34.

BARSH, G. S., SEEBURG, P. H. & GELINAS, R. E. 1983. The human growth hormone gene family: structure and evolution of the chromosomal locus. *Nucleic Acids Res,* 11**,** 3939-58.

BENTLEY, D. R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics & Development,* 16**,** 545-552.

BROECKX, B. J., HITTE, C., COOPMAN, F., VERHOEVEN, G. E., DE KEULENAER, S., DE MEESTER, E., DERRIEN, T., ALFOLDI, J., LINDBLAD-TOH, K., BOSMANS, T., GIELEN, I., VAN BREE, H., VAN RYSSEN, B., SAUNDERS, J. H., VAN NIEUWERBURGH, F. & DEFORCE, D. 2015. Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. *Sci Rep,* 5**,** 12810.

BROECKX, B. J. G., COOPMAN, F., VERHOEVEN, G. E. C., BAVEGEMS, V., DE KEULENAER, S., DE MEESTER, E., VAN NIEWERBURGH, F. & DEFORCE, D. 2014. Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). 4**,** 5597.

BROWN, J. H., JARDETZKY, T., SAPER, M. A., SAMRAOUI, B., BJORKMAN, P. J. & WILEY, D. C. 1988. A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. *Nature,* 332**,** 845-50.

BROWN, J. J., THOMSON, W., CLEGG, P., EYRE, S., KENNEDY, L. J., MATTHEWS, J., CARTER, S. & OLLIER, W. E. 2004. Polymorphisms of the equine major histocompatibility complex class II DRA locus. *Tissue Antigens,* 64**,** 173-9.

CABOCHE, S., AUDEBERT, C., LEMOINE, Y. & HOT, D. 2014. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics,* 15**,** 264.

CADUFF, M., BAUER, A., JAGANNATHAN, V. & LEEB, T. 2017. A single base deletion in the SLC45A2 gene in a Bullmastiff with oculocutaneous albinism. *Anim Genet,* 48**,** 619-621.

CARNINCI, P. 2008. Non-coding RNA transcription: turning on neighbours. *Nat Cell Biol,* 10**,** 1023-1024.

CHAISSON, M. J., HUDDLESTON, J., DENNIS, M. Y., SUDMANT, P. H., MALIG, M., HORMOZDIARI, F., ANTONACCI, F., SURTI, U., SANDSTROM, R., BOITANO, M., LANDOLIN, J. M., STAMATOYANNOPOULOS, J. A., HUNKAPILLER, M. W., KORLACH, J. & EICHLER, E. E. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature,* 517**,** 608-11.

CHILDERS, C. P., NEWKIRK, H. L., HONEYCUTT, D. A., RAMLACHAN, N., MUZNEY, D. M., SODERGREN, E., GIBBS, R. A., WEINSTOCK, G. M., WOMACK, J. E. & SKOW, L. C. 2006. Comparative analysis of the bovine MHC class IIb sequence identifies inversion breakpoints and three unexpected genes. *Anim Genet,* 37**,** 121-9.

CHIN, C. S., ALEXANDER, D. H., MARKS, P., KLAMMER, A. A., DRAKE, J., HEINER, C., CLUM, A., COPELAND, A., HUDDLESTON, J., EICHLER, E. E., TURNER, S. W. & KORLACH, J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods,* 10**,** 563-9.

CONRAD, D. F., BIRD, C., BLACKBURNE, B., LINDSAY, S., MAMANOVA, L., LEE, C., TURNER, D. J. & HURLES, M. E. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet,* 42**,** 385-91.

COSART, T., BEJA-PEREIRA, A., CHEN, S., NG, S. B., SHENDURE, J. & LUIKART, G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics,* 12**,** 347.

COX, M. L., EVANS, J. M., DAVIS, A. G., GUO, L. T., LEVY, J. R., STARR-MOSS, A. N., SALMELA, E., HYTÖNEN, M. K., LOHI, H., CAMPBELL, K. P., CLARK, L. A. & SHELTON, G. D. 2017. Exome sequencing reveals independent SGCD deletions causing limb girdle muscular dystrophy in Boston terriers. *Skeletal Muscle,* 7**,** 15.

DAY, K., SONG, J. & ABSHER, D. 2014. Targeted sequencing of large genomic regions with CATCH-Seq. *PLoS One,* 9**,** e111756.

DAY-WILLIAMS, A. G. & ZEGGINI, E. 2011. The effect of next-generation sequencing technology on complex trait research. *European Journal of Clinical Investigation,* 41**,** 561-567.

DARWIN, C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 1st ed: John Murray;.

DEBENHAM, S. L., HART, E. A., ASHURST, J. L., HOWE, K. L., QUAIL, M. A., OLLIER, W. E. R. & BINNS, M. M. 2005. Genomic sequence of the class II region of the canine MHC: comparison with the MHC of other mammalian species. *Genomics,* 85**,** 48-59.

DEHAL, P. & BOORE, J. L. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology,* 3**,** e314.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.

DIAZ, S., GIOVAMBATTISTA, G., DULOUT, F. N. & PERAL-GARCIA, P. 2001. Genetic variation of the second exon of ELA-DRB genes in Argentine Creole horses. *Anim Genet,* 32**,** 257-63.

DRISCOLL, C. A., MACDONALD, D. W. & O'BRIEN, S. J. 2009. From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences,* 106**,** 9971-9978.

DROGEMULLER, M., JAGANNATHAN, V., BECKER, D., DROGEMULLER, C., SCHELLING, C., PLASSAIS, J., KAERLE, C., DUFAURE DE CITRES, C., THOMAS, A., MULLER, E. J., WELLE, M. M., ROOSJE, P. & LEEB, T. 2014. A mutation in the FAM83G gene in dogs with hereditary footpad hyperkeratosis (HFH). *PLoS Genet,* 10**,** e1004370.

DURIG, N., JUDE, R., HOLL, H., BROOKS, S. A., LAFAYETTE, C., JAGANNATHAN, V. & LEEB, T. 2017. Whole genome sequencing reveals a novel deletion variant in the KIT gene in horses with white spotted coat colour phenotypes. *Anim Genet,* 48**,** 483-485.

EVANS, J. M., COX, M. L., HUSKA, J., LI, F., GAITERO, L., GUO, L. T., CASAL, M. L., GRANZIER, H. L., SHELTON, G. D. & CLARK, L. A. 2016. Exome sequencing reveals a nebulin nonsense mutation in a dog model of nemaline myopathy. *Mamm Genome,* 27**,** 495-502.

EWING, B. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res,* 8**,** 186-94.

EZKURDIA, I., JUAN, D., RODRIGUEZ, J. M., FRANKISH, A., DIEKHANS, M., HARROW, J., VAZQUEZ, J., VALENCIA, A. & TRESS, M. L. 2014. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics,* 23**,** 5866-5878.

FARNHAM, P. J. 2009. Insights from genomic profiling of transcription factors. *Nat Rev Genet,* 10**,** 605-616.

FAULKNER, G. J., KIMURA, Y., DAUB, C. O., WANI, S., PLESSY, C., IRVINE, K. M., SCHRODER, K., CLOONAN, N., STEPTOE, A. L., LASSMANN, T., WAKI, K., HORNIG, N., ARAKAWA, T., TAKAHASHI, H., KAWAI, J., FORREST, A. R. R., SUZUKI, H., HAYASHIZAKI, Y., HUME, D. A., ORLANDO, V., GRIMMOND, S. M. & CARNINCI, P. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet,* 41**,** 563-571.

FLAJNIK, M. F. & KASAHARA, M. 2001. Comparative Genomics of the MHC: Glimpses into the Evolution of the Adaptive Immune System. *Immunity,* 15**,** 351-362.

FLUSBERG, B. A., WEBSTER, D., LEE, J., TRAVERS, K., OLIVARES, E., CLARK, T. A., KORLACH, J. & TURNER, S. W. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods,* 7**,** 461-465.

FRANK, J. A., PAN, Y., TOOMING-KLUNDERUD, A., EIJSINK, V. G. H., MCHARDY, A. C., NEDERBRAGT, A. J. & POPE, P. B. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports,* 6**,** 25373.

FRASER, D. G. & BAILEY, E. 1996. Demonstration of threeDRB locl in a domestic horse family. *Immunogenetics,* 44**,** 441-445.

FRASER, D. G. & BAILEY, E. 1998. Polymorphism and multiple loci for the horse DQA gene. *Immunogenetics,* 47**,** 487-490.

FRISCHKNECHT, M., NIEHOF-OELLERS, H., JAGANNATHAN, V., OWCZAREK-LIPSKA, M., DROGEMULLER, C., DIETSCHI, E., DOLF, G., TELLHELM, B., LANG, J., TIIRA, K., LOHI, H. & LEEB, T. 2013. A COL11A2 mutation in Labrador retrievers with mild disproportionate dwarfism. *PLoS One,* 8**,** e60149.

FRITZ, K. L., KAESE, H. J., VALBERG, S. J., HENDRICKSON, J. A., RENDAHL, A. K., BELLONE, R. R., DYNES, K. M., WAGNER, M. L., LUCIO, M. A., CUOMO, F. M., BRINKMEYER-LANGFORD, C. L., SKOW, L. C., MICKELSON, J. R., RUTHERFORD, M. S. & MCCUE, M. E. 2014. Genetic risk factors for insidious equine recurrent uveitis in Appaloosa horses. *Anim Genet,* 45**,** 392-9.

FUJII, J., OTSU, K., ZORZATO, F., DE LEON, S., KHANNA, V. K., WEILER, J. E., O'BRIEN, P. J. & MACLENNAN, D. H. 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science,* 253**,** 448-51.

FUJITA, T. & FUJII, H. 2013. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochemical and Biophysical Research Communications,* 439**,** 132-136.

FUSCO, G. & MINELLI, A. 2010. Phenotypic plasticity in development and evolution: facts and concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 547-556.

GABRIELI, T., SHARIM, H., MICHAELI, Y. & EBENSTEIN, Y. 2017. Cas9-Assisted Targeting of CHromosome segments (CATCH) for targeted nanopore sequencing and optical genome mapping. *bioRxiv*.

GAO, J., LIU, K., LIU, H., BLAIR, H. T., LI, G., CHEN, C., TAN, P. & MA, R. Z. 2010. A complete DNA sequence map of the ovine major histocompatibility complex. *BMC Genomics,* 11**,** 466.

GHURYE, J., POP, M., KOREN, S., BICKHART, D. & CHIN, C.-S. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genomics,* 18**,** 527.

GODDARD, M. E. & HAYES, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet,* 10**,** 381-391.

GOFFEAU, A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J. D., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H. & OLIVER, S. G. 1996. Life with 6000 genes. *Science,* 274**,** 546, 563-7.

GOLDBERG, M. L., SHEEN, J. Y., GEHRING, W. J. & GREEN, M. M. 1983. Unequal crossing-over associated with asymmetrical synapsis between nomadic elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A,* 80**,** 5017-21.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-351.

GORDON, D., HUDDLESTON, J., CHAISSON, M. J., HILL, C. M., KRONENBERG, Z. N., MUNSON, K. M., MALIG, M., RAJA, A., FIDDES, I., HILLIER, L. W., DUNN, C., BAKER, C., ARMSTRONG, J., DIEKHANS, M., PATEN, B., SHENDURE, J., WILSON, R. K., HAUSSLER, D., CHIN, C. S. & EICHLER, E. E. 2016. Long-read sequence assembly of the gorilla genome. *Science,* 352**,** aae0344.

GORER, P. A., LYMAN, S. & SNELL, G. D. 1948. Studies on the Genetic and Antigenic Basis of Tumour Transplantation. Linkage between a Histocompatibility Gene and 'Fused' in Mice. *Proceedings of the Royal Society of London B: Biological Sciences,* 135**,** 499-505.

GRAY, Y. H. M. 2000. It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. *Trends in Genetics,* 16**,** 461-468.

GREGORY, T. R. 2005. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet,* 6**,** 699-708.

GUSTAFSON, A. L., TALLMADGE, R. L., RAMLACHAN, N., MILLER, D., BIRD, H., ANTCZAK, D. F., RAUDSEPP, T., CHOWDHARY, B. P. & SKOW, L. C. 2003. An ordered BAC contig map of the equine major histocompatibility complex. *Cytogenetic and Genome Research,* 102**,** 189-195.

GUSTAFSSON, K. & ANDERSSON, L. 1994. Structure and polymorphism of horse MHC class II DRB genes: convergent evolution in the antigen binding site. *Immunogenetics,* 39**,** 355-8.

HANS, J. B., BERGL, R. A. & VIGILANT, L. 2017. Gorilla MHC class I gene and sequence variation in a comparative context. *Immunogenetics,* 69**,** 303-323.

HEATHER, J. M. & CHAIN, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

HILLIER, L. W., COULSON, A., MURRAY, J. I., BAO, Z., SULSTON, J. E. & WATERSTON, R. H. 2005. Genomics in C. elegans: so many genes, such a little worm. *Genome Res,* 15**,** 1651-60.

HOEPPNER, M. P., LUNDQUIST, A., PIRUN, M., MEADOWS, J. R., ZAMANI, N., JOHNSON, J., SUNDSTROM, G., COOK, A., FITZGERALD, M. G., SWOFFORD, R., MAUCELI, E., MOGHADAM, B. T., GREKA, A., ALFOLDI, J., ABOUELLEIL, A., AFTUCK, L., BESSETTE, D., BERLIN, A., BROWN, A., GEARIN, G., LUI, A., MACDONALD, J. P., PRIEST, M., SHEA, T., TURNER-MAIER, J., ZIMMER, A., LANDER, E. S., DI PALMA, F., LINDBLAD-TOH, K. & GRABHERR, M. G. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One,* 9**,** e91172.

HOGEWEG, P. 2011. The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology,* 7**,** e1002021.

HOLLING, T. M., SCHOOTEN, E. & VAN DEN ELSEN, P. J. 2004. Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. *Hum Immunol,* 65**,** 282-90.

HORIN, P. & MATIASOVIC, J. 2002. A second locus and new alleles in the major histocompatibility complex class II (ELA-DQB) region in the horse. *Anim Genet,* 33**,** 196-200.

HUDDLESTON, J., CHAISSON, M. J. P., STEINBERG, K. M., WARREN, W., HOEKZEMA, K., GORDON, D., GRAVES-LINDSAY, T. A., MUNSON, K. M., KRONENBERG, Z. N., VIVES, L., PELUSO, P., BOITANO, M., CHIN, C. S., KORLACH, J., WILSON, R. K. &

EICHLER, E. E. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res,* 27**,** 677-685.

HUDDLESTON, J., RANADE, S., MALIG, M., ANTONACCI, F., CHAISSON, M., HON, L., SUDMANT, P. H., GRAVES, T. A., ALKAN, C., DENNIS, M. Y., WILSON, R. K., TURNER, S. W., KORLACH, J. & EICHLER, E. E. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res,* 24**,** 688-96.

HUGHES, A. L., OTA, T. & NEI, M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Molecular Biology and Evolution,* 7**,** 515-524.

HUNT, M., NEWBOLD, C., BERRIMAN, M. & OTTO, T. D. 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome Biology,* 15**,** R42.

HURT, P., WALTER, L., SUDBRAK, R., KLAGES, S., MÜLLER, I., SHIINA, T., INOKO, H., LEHRACH, H., GÜNTHER, E., REINHARDT, R. & HIMMELBAUER, H. 2004. The Genomic Sequence and Comparative Analysis of the Rat Major Histocompatibility Complex. *Genome Research,* 14**,** 631-639.

HWANG, S., KIM, E., LEE, I. & MARCOTTE, E. M. 2015. Systematic comparison of variant calling pipelines using gold standard personal exome variants. 5**,** 17875.

YUHKI, N., BECK, T., STEPHENS, R., NEELAM, B. & O'BRIEN, S. J. 2007. Comparative genomic structure of human, dog, and cat MHC: HLA, DLA, and FLA. *J Hered,* 98**,** 390-9.

JAGANNATHAN, V., BANNOEHR, J., PLATTET, P., HAUSWIRTH, R., DROGEMULLER, C., DROGEMULLER, M., WIENER, D. J., DOHERR, M., OWCZAREK-LIPSKA, M., GALICHET, A., WELLE, M. M., TENGVALL, K., BERGVALL, K., LOHI, H., RUFENACHT, S., LINEK, M., PARADIS, M., MULLER, E. J., ROOSJE, P. & LEEB, T. 2013. A mutation in the SUV39H2 gene in Labrador Retrievers with hereditary nasal parakeratosis (HNPK) provides insights into the epigenetics of keratinocyte differentiation. *PLoS Genet,* 9**,** e1003848.

JAIN, M., FIDDES, I. T., MIGA, K. H., OLSEN, H. E., PATEN, B. & AKESON, M. 2015. Improved data analysis for the MinION nanopore sequencer. *Nat Meth,* 12**,** 351-356.

JAIN, M., KOREN, S., QUICK, J., RAND, A. C., SASANI, T. A., TYSON, J. R., BEGGS, A. D., DILTHEY, A. T., FIDDES, I. T., MALLA, S., MARRIOTT, H., MIGA, K. H., NIETO, T., GRADY, J., OLSEN, H. E., PEDERSEN, B. S., RHIE, A., RICHARDSON, H., QUINLAN, A., SNUTCH, T. P., TEE, L., PATEN, B., PHILLIPPY, A. M., SIMPSON, J. T., LOMAN, N. J. & LOOSE, M. 2017. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*.

JIAO, Y., PELUSO, P., SHI, J., LIANG, T., STITZER, M. C., WANG, B., CAMPBELL, M. S., STEIN, J. C., WEI, X., CHIN, C.-S., GUILL, K., REGULSKI, M., KUMARI, S., OLSON, A., GENT, J., SCHNEIDER, K. L., WOLFGRUBER, T. K., MAY, M. R., SPRINGER, N. M., ANTONIOU, E., MCCOMBIE, W. R., PRESTING, G. G., MCMULLEN, M., ROSS-IBARRA, J., DAWE, R. K., HASTIE, A., RANK, D. R. & WARE, D. 2017. Improved maize reference genome with single-molecule technologies. *Nature,* advance online publication.

KAESSMANN, H., VINCKENBOSCH, N. & LONG, M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet,* 10**,** 19-31.

KAMATH, P. L. & GETZ, W. M. 2011. Adaptive molecular evolution of the Major Histocompatibility Complex genes, DRA and DQA, in the genus Equus. *BMC Evol Biol,* 11**,** 128.

KARLSSON, E. K. & LINDBLAD-TOH, K. 2008. Leader of the pack: gene mapping in dogs and other model organisms. *Nat Rev Genet,* 9.

KELLEY, J., WALTER, L. & TROWSDALE, J. 2005. Comparative genomics of major histocompatibility complexes. *Immunogenetics,* 56**,** 683-95.

KIMCHI-SARFATY, C., OH, J. M., KIM, I. W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V. & GOTTESMAN, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science,* 315**,** 525-8.

KLUMPLEROVA, M., VYCHODILOVA, L., BOBROVA, O., CVANOVA, M., FUTAS, J., JANOVA, E., VYSKOCIL, M., VRTKOVA, I., PUTNOVA, L., DUSEK, L., MARTI, E. & HORIN, P. 2013. Major histocompatibility complex and other allergy-related candidate genes associated with insect bite hypersensitivity in Icelandic horses. *Mol Biol Rep,* 40**,** 3333-40.

LAIRD, P. W. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet,* 11**,** 191-203.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, Y., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LEVENE, M. J., KORLACH, J., TURNER, S. W., FOQUET, M., CRAIGHEAD, H. G. & WEBB, W. W. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science,* 299**,** 682-6.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H., RUAN, J. & DURBIN, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research,* 18**,** 1851-1858.

LI, W. & FREUDENBERG, J. 2014. Mappability and read length. *Frontiers in Genetics,* 5**,** 381.

LYNCH, VINCENT J., NNAMANI, MAURIS C., KAPUSTA, A., BRAYER, K., PLAZA, SILVIA L., MAZUR, ERIK C., EMERA, D., SHEIKH, SHEHZAD Z., GRÜTZNER, F., BAUERSACHS, S., GRAF, A., YOUNG, STEVEN L., LIEB, JASON D., DEMAYO, FRANCESCO J., FESCHOTTE, C. & WAGNER, GÜNTER P. 2015. Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Reports,* 10**,** 551-561.

LINDBLAD-TOH, K., GARBER, M., ZUK, O., LIN, M. F., PARKER, B. J., WASHIETL, S., KHERADPOUR, P., ERNST, J., JORDAN, G., MAUCELI, E., WARD, L. D., LOWE, C. B., HOLLOWAY, A. K., CLAMP, M., GNERRE, S., ALFOLDI, J., BEAL, K., CHANG, J., CLAWSON, H., CUFF, J., DI PALMA, F., FITZGERALD, S., FLICEK, P., GUTTMAN, M., HUBISZ, M. J., JAFFE, D. B., JUNGREIS, I., KENT, W. J., KOSTKA, D., LARA, M., MARTINS, A. L., MASSINGHAM, T., MOLTKE, I., RANEY, B. J., RASMUSSEN, M. D., ROBINSON, J., STARK, A., VILELLA, A. J., WEN, J., XIE, X., ZODY, M. C., WORLEY, K. C., KOVAR, C. L., MUZNY, D. M., GIBBS, R. A., WARREN, W. C., MARDIS, E. R., WEINSTOCK, G. M., WILSON, R. K., BIRNEY, E., MARGULIES, E. H., HERRERO, J., GREEN, E. D., HAUSSLER, D., SIEPEL, A., GOLDMAN, N., POLLARD, K. S., PEDERSEN, J. S., LANDER, E. S. & KELLIS, M. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature,* 478**,** 476-482.

LUPSKI, J. R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics,* 14**,** 417-422.

MAGI, A., SEMERARO, R., MINGRINO, A., GIUSTI, B. & D'AURIZIO, R. 2017. Nanopore sequencing data analysis: state of the art, applications and challenges. *Brief Bioinform*.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y.-J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L. I., JARVIE, T. P., JIRAGE, K. B., KIM, J.-B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

MARTIN, J. A. & WANG, Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet,* 12**,** 671-682.

MCCLINTOCK, B. 1951. Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol,* 16**,** 13-47.

MCCLURE, M. C., BICKHART, D., NULL, D., VANRADEN, P., XU, L., WIGGANS, G., LIU, G., SCHROEDER, S., GLASSCOCK, J., ARMSTRONG, J., COLE, J. B., VAN TASSELL, C. P. & SONSTEGARD, T. S. 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. *PLoS One,* 9**,** e92769.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MEADOWS, J. R. S. & LINDBLAD-TOH, K. 2017. Dissecting evolution and disease using comparative vertebrate genomics. *Nat Rev Genet,* 18**,** 624-636.

MERKER, J. D., WENGER, A. M., SNEDDON, T., GROVE, M., ZAPPALA, Z., FRESARD, L., WAGGOTT, D., UTIRAMERUR, S., HOU, Y., SMITH, K. S., MONTGOMERY, S. B., WHEELER, M., BUCHAN, J. G., LAMBERT, C. C., ENG, K. S., HICKEY, L., KORLACH, J., FORD, J. & ASHLEY, E. A. 2017. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*.

MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95**,** 315-327.

MUHLETHALER-MOTTET, A., OTTEN, L. A., STEIMLE, V. & MACH, B. 1997. Expression of MHC class II molecules in different cellular and functional compartments is controlled by differential usage of multiple promoters of the transactivator CIITA. *The EMBO Journal,* 16**,** 2851-2860.

NICHOLAS, F. W. 2005. Animal breeding and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 360**,** 1529-1536.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics,* 12**,** 443-451.

NYREN, P., PETTERSSON, B. & UHLEN, M. 1993. Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry,* 208**,** 171-175.

OHNO, S. 1970. *Evolution by Gene Duplication* Springer Verlag, Berlin.

OSOEGAWA, K., WOON, P. Y., ZHAO, B., FRENGEN, E., TATENO, M., CATANESE, J. J. & DE JONG, P. J. 1998. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics,* 52**,** 1-8.

OWCZAREK-LIPSKA, M., JAGANNATHAN, V., DROGEMULLER, C., LUTZ, S., GLANEMANN, B., LEEB, T. & KOOK, P. H. 2013. A frameshift mutation in the cubilin gene (CUBN) in Border Collies with Imerslund-Grasbeck syndrome (selective cobalamin malabsorption). *PLoS One,* 8**,** e61144.

OZAKI, K., OHNISHI, Y., IIDA, A., SEKINE, A., YAMADA, R., TSUNODA, T., SATO, H., SATO, H., HORI, M., NAKAMURA, Y. & TANAKA, T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet,* 32**,** 650-4.

PARK, P. J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet,* 10**,** 669-80.

PETERSON, D. G., SCHULZE, S. R., SCIARA, E. B., LEE, S. A., BOWERS, J. E., NAGEL, A., JIANG, N., TIBBITTS, D. C., WESSLER, S. R. & PATERSON, A. H. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res,* 12**,** 795-807.

PIERTNEY, S. B. & OLIVER, M. K. 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb),* 96**,** 7-21.

REESE, M. G., HARTZELL, G., HARRIS, N. L., OHLER, U., ABRIL, J. F. & LEWIS, S. E. 2000. Genome annotation assessment in Drosophila melanogaster. *Genome Res,* 10**,** 483-501.

RICHARD, G. F., KERREST, A. & DUJON, B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev,* 72**,** 686-727.

RICHTERICH, P. 1998. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res,* 8**,** 251-9.

ROE, D., VIERRA-GREEN, C., PYO, C. W., ENG, K., HALL, R., KUANG, R., SPELLMAN, S., RANADE, S., GERAGHTY, D. E. & MAIERS, M. 2017. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun*.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J., EDWARDS, M., HOON, J., SIMONS, J. F., MARRAN, D., MYERS, J. W., DAVIDSON, J. F., BRANTING, A., NOBILE, J. R., PUC, B. P., LIGHT, D., CLARK, T. A., HUBER, M., BRANCIFORTE, J. T., STONER, I. B., CAWLEY, S. E., LYONS, M., FU, Y., HOMER, N., SEDOVA, M., MIAO, X., REED, B., SABINA, J., FEIERSTEIN, E., SCHORN, M., ALANJARY, M., DIMALANTA, E., DRESSMAN, D., KASINSKAS, R., SOKOLSKY, T., FIDANZA, J. A., NAMSARAEV, E., MCKERNAN, K. J., WILLIAMS, A., ROTH, G. T. & BUSTILLO, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475**,** 348-52.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-5467.

SANTAGOSTINO, M., KHORIAULI, L., GAMBA, R., BONUGLIA, M., KLIPSTEIN, O., PIRAS, F. M., VELLA, F., RUSSO, A., BADIALE, C., MAZZAGATTI, A., RAIMONDI, E., NERGADZE, S. G. & GIULOTTO, E. 2015. Genome-wide evolutionary and functional analysis of the Equine Repetitive Element 1: an insertion in the myostatin promoter affects gene expression. *BMC Genet,* 16**,** 126.

SCHBATH, S., MARTIN, V., ZYTNICKI, M., FAYOLLE, J., LOUX, V. & GIBRAT, J.-F. 2012. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology,* 19**,** 796-813.

SCHURINK, A., WOLC, A., DUCRO, B. J., FRANKENA, K., GARRICK, D. J., DEKKERS, J. C. & VAN ARENDONK, J. A. 2012. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. *Genetics Selection Evolution,* 44**,** 1-12.

SOHN, J. I. & NAM, J. W. 2016. The present and future of de novo whole-genome assembly. *Brief Bioinform*.

STADEN, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research,* 6(7)**,** 2601-2610.

STAIGER, E. A., TSENG, C. T., MILLER, D., CASSANO, J. M., NASIR, L., GARRICK, D., BROOKS, S. A. & ANTCZAK, D. F. 2016. Host genetic influence on papillomavirus-induced tumors in the horse. *Int J Cancer*.

STANKE, M. & WAACK, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics,* 19 Suppl 2**,** ii215-25.

STONEKING, M. 2001. Single nucleotide polymorphisms: From the evolutionary past. *Nature,* 409**,** 821-822.

TAFT, R. J., GLAZOV, E. A., CLOONAN, N., SIMONS, C., STEPHEN, S., FAULKNER, G. J., LASSMANN, T., FORREST, A. R. R., GRIMMOND, S. M., SCHRODER, K., IRVINE, K., ARAKAWA, T., NAKAMURA, M., KUBOSAKI, A., HAYASHIDA, K., KAWAZU, C., MURATA, M., NISHIYORI, H., FUKUDA, S., KAWAI, J., DAUB, C. O., HUME, D. A., SUZUKI, H., ORLANDO, V., CARNINCI, P., HAYASHIZAKI, Y. & MATTICK, J. S. 2009. Tiny RNAs associated with transcription start sites in animals. *Nat Genet,* 41**,** 572-578.

TAKAHASHI, K., ROONEY, A. P. & NEI, M. 2000. Origins and divergence times of mammalian class II MHC gene clusters. *J Hered,* 91**,** 198-204.

TALLMADGE, R. L., LEAR, T. L. & ANTCZAK, D. F. 2005. Genomic characterization of MHC class I genes of the horse. *Immunogenetics,* 57**,** 763-74.

TELENTI, A., PIERCE, L. C. T., BIGGS, W. H., DI IULIO, J., WONG, E. H. M., FABANI, M. M., KIRKNESS, E. F., MOUSTAFA, A., SHAH, N., XIE, C., BREWERTON, S. C., BULSARA, N., GARNER, C., METZKER, G., SANDOVAL, E., PERKINS, B. A., OCH, F. J., TURPAZ, Y. & VENTER, J. C. 2016. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences,* 113**,** 11901-11906.

THE GENOMES PROJECT, C. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

THE MHC SEQUENCING CONSORTIUM 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature,* 401**,** 921-923.

TING, J. P. & TROWSDALE, J. 2002. Genetic control of MHC class II expression. *Cell,* 109 Suppl**,** S21-33.

TSUBOI, M., WATANABE, M., NIBE, K., YOSHIMI, N., KATO, A., SAKAGUCHI, M., YAMATO, O., TANAKA, M., KUWAMURA, M., KUSHIDA, K., ISHIKURA, T., HARADA, T., CHAMBERS, J. K., SUGANO, S., UCHIDA, K. & NAKAYAMA, H. 2017. Identification of the PLA2G6 c.1579G>A Missense Mutation in Papillon Dog Neuroaxonal Dystrophy Using Whole Exome Sequencing Analysis. *PLoS ONE,* 12**,** e0169002.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-51.

VIJ, S., KUHL, H., KUZNETSOVA, I. S., KOMISSAROV, A., YURCHENKO, A. A., VAN HEUSDEN, P., SINGH, S., THEVASAGAYAM, N. M., PRAKKI, S. R., PURUSHOTHAMAN, K., SAJU, J. M., JIANG, J., MBANDI, S. K., JONAS, M., HIN YAN TONG, A., MWANGI, S., LAU, D., NGOH, S. Y., LIEW, W. C., SHEN, X., HON, L. S., DRAKE, J. P., BOITANO, M., HALL, R., CHIN, C. S., LACHUMANAN, R., KORLACH, J., TRIFONOV, V., KABILOV, M., TUPIKIN, A., GREEN, D., MOXON, S., GARVIN, T., SEDLAZECK, F. J., VURTURE, G. W., GOPALAPILLAI, G., KUMAR KATNENI, V., NOBLE, T. H., SCARIA, V., SIVASUBBU, S., JERRY, D. R., O'BRIEN, S. J., SCHATZ, M. C., DALMAY, T., TURNER, S. W., LOK, S., CHRISTOFFELS, A. & ORBAN, L. 2016. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genet,* 12**,** e1005954.

VIĻUMA, A., MIKKO, S., HAHN, D., SKOW, L., ANDERSSON, G. & BERGSTRÖM, T. F. 2017. Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology. *Scientific Reports,* 7**,** 45518.

WADE, C. M., GIULOTTO, E., SIGURDSSON, S., ZOLI, M., GNERRE, S., IMSLAND, F., LEAR, T. L., ADELSON, D. L., BAILEY, E., BELLONE, R. R., BLOCKER, H., DISTL, O., EDGAR, R. C., GARBER, M., LEEB, T., MAUCELI, E., MACLEOD, J. N., PENEDO, M. C., RAISON, J. M., SHARPE, T., VOGEL, J., ANDERSSON, L., ANTCZAK, D. F., BIAGI, T., BINNS, M. M., CHOWDHARY, B. P., COLEMAN, S. J., DELLA VALLE, G., FRYC, S., GUERIN, G., HASEGAWA, T., HILL, E. W., JURKA, J., KIIALAINEN, A., LINDGREN, G., LIU, J., MAGNANI, E., MICKELSON, J. R., MURRAY, J., NERGADZE, S. G., ONOFRIO, R., PEDRONI, S., PIRAS, M. F., RAUDSEPP, T., ROCCHI, M., ROED, K. H., RYDER, O. A., SEARLE, S., SKOW, L., SWINBURNE, J. E., SYVANEN, A. C., TOZAKI, T., VALBERG, S. J., VAUDIN, M., WHITE, J. R., ZODY, M. C., BROAD INSTITUTE GENOME SEQUENCING, P., BROAD INSTITUTE WHOLE GENOME ASSEMBLY, T., LANDER, E. S. & LINDBLAD-TOH, K. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science,* 326**,** 865-7.

WAN, Q. H., ZENG, C. J., NI, X. W., PAN, H. J. & FANG, S. G. 2009. Giant panda genomic data provide insight into the birth-and-death process of mammalian major histocompatibility complex class II genes. *PLoS One,* 4**,** e4147.

WANG, D. G., FAN, J.-B., SIAO, C.-J., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLYAK, L., STEIN, L., HSIE, L., TOPALOGLOU, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M. S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T. J., LIPSHUTZ, R., CHEE, M. & LANDER, E. S. 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science,* 280**,** 1077-1082.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research,* 38**,** e164-e164.

WANG, Z., CHEN, Y. & LI, Y. 2004. A Brief Review of Computational Gene Prediction Methods. *Genomics, Proteomics & Bioinformatics,* 2**,** 216-221.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WATSON, J. D. & CRICK, F. H. C. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature,* 171**,** 737-738.

WOLFE, K. H. & LI, W. H. 2003. Molecular evolution meets the genomics revolution. *Nat Genet,* 33 Suppl**,** 255-65.

ZHANG, H. 2003. Evolution by gene duplication: an update. *Trends in Ecology and Evolution,* 18**,** 292-298.