University of Veterinary Medicine Hannover

Exploring genomes of domestic animals using Next Generation Sequencing

Habilitation
to obtain the *Venia Legendi* at the University of Veterinary Medicine Hannover

submitted by
Dr. med. vet. Julia Anna Petra Metzger

Hannover 2018

Day of non-public scientific debate: 05.11.2018

Parts of this study were financially supported by the German Research Foundation (DFG; ME 4517/1-1).

To my family

Table of contents

1 List of abbreviations

ABT1 .... *activator of basal transcription 1*
ACAN .... *aggrecan*
ADAMTS17 .... *disintegrin-like and metalloprotease with thrombospondin type 1 motif 17*
CFA .... canis familiaris autosome
CFTR .... *cystic fibrosis transmembrane conductance regulator*
CNVs .... copy number variants
CT .... computer tomography
DIAPH3 .... *diaphanous related formin 3*
DNA .... deoxyribonucleic acid
dNTPs .... desoxy nucleoside triphosphate
EBV-PAT .... breeding values of the paternal component of the pregnancy rate per estrus
ECA .... equus caballus autosome
EM .... expectation-maximization
FBXO43 .... *F-box protein 43*
FOXP1 .... *forkhead box P1*
GH1 .... *growth hormone 1*
GWAS .... genome-wide association study
HAS2 .... *hyaluronic acid synthase 2*
HMGA2 .... *high mobility group AT-hook 2*
indel .... insertion/deletion
KITLG .... *KIT ligand*
KRT1 .... *keratin gene 1*
KRT25 .... *keratin 25*
KRT27 .... *keratin 27*
KRTAP16 .... *keratin associated protein 16*
LS .... *Lundehund syndrome*
MAF .... minor allele frequency
MTBP .... *MDM2 binding protein*
ND3 .... *mitochondrially encoded NADH dehydrogenase 3*
ND6 .... *mitochondrially encoded NADH dehydrogenase 3*
NGS .... Next Generation Sequencing
NOD1 .... *nucleotide-binding oligomerization domain containing 1*
NOTCH1 .... *neurogenic locus notch homolog protein 1*
nsSNPs .... non-synonymous single nucleotide polymorphisms
OSTN .... *osteocrin*
OVGP1 .... *oviductal glycoprotein 1*
PKD1 .... *polycystic kidney disease-associated protein*
PLE .... *protein-losing enteropathy*
ROH .... runs of homozygosity
SEC31B .... *SEC31-like protein 2*
SEM .... scanning electron microscopy

SNPs .... single nucleotide polymorphisms
SP6 .... *transcription factor SP6*
SPAID .... Shar-Pei Autoinflammatory Disease
SPATA31E1 .... *SPATA31 subfamily E member 1*
sSNPs .... synonymous single nucleotide polymorphisms
TGFBR3 .... *transforming growth factor beta receptor 3*
Tp53 .... *cellular tumor antigen p53*
TSSK6 .... *testis specific serine kinase 6*
YES1 .... *V-Yes-1 yamaguchi sarcoma viral oncogene homolog 1*

2 List of publications which are parts of this work

Publication 1 (appendix 12.1):

Metzger J, Tonda R, Beltran S, Agueda L, Gut M, Distl O: Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. BMC Genomics. 2014;15:562.

Publication 2 (appendix 12.2):

Metzger J, Karwath M, Tonda R, Beltran S, Agueda L, Gut M, Gut IG, Distl O: Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics. 2015;16:764.

Publication 3 (appendix 12.3):

Schrimpf R, Gottschalk M, **Metzger J**, Martinsson G, Sieme H, Distl O: Screening of whole genome sequences identified high-impact variants for stallion fertility. BMC Genomics. 2016;17:288.

Publication 4 (appendix 12.4):

Metzger J, Rau J, Conn LB, Lindgren G, Distl O: Synergism of four key regulators determines extremely small body size in horses. BMC Genomics, 2018; 19(1):492, doi: 10.1186/s12864-018-4877-5.

Publication 5 (appendix 12.5):

Metzger J, Pfahler S, Distl O: Variant detection and runs of homozygosity in next generation sequencing data elucidate the genetic background of Lundehund syndrome. BMC Genomics. 2016;17:535.

Publication 6 (appendix 12.6):

Metzger J, Distl O: A study of Shar-Pei dogs refutes association of the 'meatmouth' duplication near HAS2 with Familial Shar-Pei Fever. Animal Genetics. 2014;45(5):763-764.

Publication 7 (appendix 12.7):

Metzger J, Nolte A, Uhde AK, Hewicker-Trautwein M, Distl O: Whole genome sequencing identifies missense mutation in MTBP in Shar-Pei affected with Autoinflammatory Disease (SPAID). BMC Genomics. 2017;18(1):348.

Publication 8 (appendix 12.8):

Metzger J, Gast AC, Schrimpf R, Rau J, Eikelberg D, Beineke A, Hellige M, Distl O: Whole-genome sequencing reveals a potential causal mutation for dwarfism in the Miniature Shetland pony. Mammalian Genome. 2017;28(3-4):143-151.

Publication 9 (appendix 12.9):

Thomer A, Gottschalk M, Christmann A, Naccache F, Jung K, Hewicker-Trautwein M, Distl O, **Metzger J**: An epistatic effect of KRT25 on SP6 is involved in curly coat in horses. Scientific Reports. 2018;8:6374, doi: 10.1038/s41598-018-24865-3.

Publication 10 (appendix 12.10):

Braun M, Reinartz S, Heppelmann M, Rehage J, Sürie C, Distl O, **Metzger J**: Curly coat caused by a keratin 27 variant was transmitted from Fleckvieh into German Angus. Animal Genetics. 2018, doi: 10.1111/age.12669

3 Preamble

The exploration of genomes harboring billions of genetic variants and their functional impact in mammals is one of the challenges of our time. Strong technical advances have changed the landscape of genetics and shifted the focus on genomics as a widespread field in domestic animal research.

Since the first sequencing of a whole genome of a bacteriophage in 1977 by Frederick Sanger and subsequent development of massively parallel sequencing thirty years later, fast and comparatively low cost sequencing of whole genomes using Next Generation Sequencing (NGS) technology led to a breakthrough in the field of genomics. NGS analyses have contributed to the identification of genetic variants ranging from coding or non-coding single nucleotide polymorphisms to insertions, deletions and structural variations. Thus, the previous single candidate gene-sequencing approaches were superseded by whole genome and transcriptome studies. The success of high-throughput sequencing has led to a continuous development and improvement of sequencing technology, particularly with regard to the length of sequence reads. A third generation of sequencing has been established, which enables single molecule sequencing and has brought great progress for the elucidation of novel genomes. These advances in the sequencing field have raised questions not only about gene and variant identification but also about organization and gene regulation in the genome. The ENCODE (Encyclopedia of DNA Elements) project for human functional genome annotation represents a pioneer for further delineation of all functional elements encoded in the genome by the implementation of high-throughput sequencing technologies.

The aim of this habilitation thesis is to demonstrate the powerful use of whole genome data derived from NGS for versatile analyses in domestic animals. It elucidates the consequence of targeted human selection for performance or conformation traits. Selection signatures and population specific variants are investigated for their potential effects resulting in desirable traits like small body size and fertility or in malformations and diseases.

4 Introduction

The fundamental basis for the formation, development and maintenance of life is the genome in its entirety. Information is packed on various chromosomes and the mitochondrial DNA composed of systematically arranged nucleotides (Mullinger & Johnson 1980). Each of these nucleotides contribute to the built of a DNA-sequence that determines function and properties of all living organisms. The base sequence of these nucleotides varies widely not only among different species and breeds but also among individuals. Whole genome sequencing performed by high throughput Next Generation Sequencing (NGS) enables to detect these sequence variations of different types including single nucleotide polymorphisms (SNPs), insertions and deletions as well as further structural variations (DePristo et al. 2011). They represent the key for understanding phenotypic variations in domestic animal genomes, especially those introduced by natural or targeted artificial selection.

Sequencing technologies- the next generation

High-throughput sequencing has become the most important and indispensable technology for genome analyses. It represents a powerful massively parallel sequencing technology allowing cost-effective and fast detection of thousands of variants at once by producing millions of reads in one run (Wheeler et al. 2008). For preparation, DNA needs to be fragmented physically or enzymatically and adaptor-ligated into libraries of about 500-900 nucleotides (Knierim et al. 2011). After attachment of these libraries onto a glass plate or on beads, sequencing-by-synthesis reactions are performed (Ju et al. 2006; Bentley et al. 2008). In the most common sequencing platforms, the Illumina sequencers, single fluorescence-labeled nucleotides (3'-O-azidomethyl-dNTPs) incorporate with the template at the complementary base temporarily terminating the polymerization during fluorescent imaging (Ju et al. 2006; Bentley et al. 2008). Finally, cleavage of this 3' block allows the incorporation of the next nucleotide. As fluorescent signals from a single event are not sufficient for detection in the four-color imaging system, a clonal amplification for all templates is performed in advance (Ju et al. 2006; Metzker 2010). Other platforms like Ion Torrent use a semiconductor technology measuring pH changes instead of fluorescence signals during DNA extension and discriminate bases by a voltage signal (Rothberg et al. 2011).

Data derived from NGS runs comprise millions of reads that have to be mapped to a reference genome or de novo without a reference for further evaluation (McKenna et al. 2010; DePristo et al. 2011). The use of different tools for further processing of reads is dependent on the input and objectives of NGS projects (Fig. 1). In particular, paired end approaches sequencing both ends of a fragment were shown to be advantageous due to particular low error rates (Dohm et al. 2008). Whole genome data derived from sequencing-by-synthesis reactions usually undergo variant calling pipelines, which comprise quality control, alignment, duplicate marking, variant calling and variant effect prediction (Pabinger et al. 2014). Reads are investigated for their sequence length and trimmed on basis of sequence quality parameters in order to improve mean read quality (Andrews 2010; Schmieder & Edwards 2011). After mapping, duplicates, that means reads pairs with the same unclipped alignment start, are marked with tags, which can be identified by downstream tools (Li & Durbin 2009; Li et al. 2009a). In a next step, variants are commonly called by the genome analysis tool kit (GATK) (McKenna et al. 2010). The use of the GATK Unified Genotyper or Haplotype Caller was controversially discussed, although more recent studies emphasized the high specificity and reliability of Haplotype Caller for indel detection (Ghoneim et al. 2014; Belkadi et al. 2015). Variant effect estimation is finally performed for functional classification of the individual SNPs/SNVs and indels (Cingolani et al. 2012).

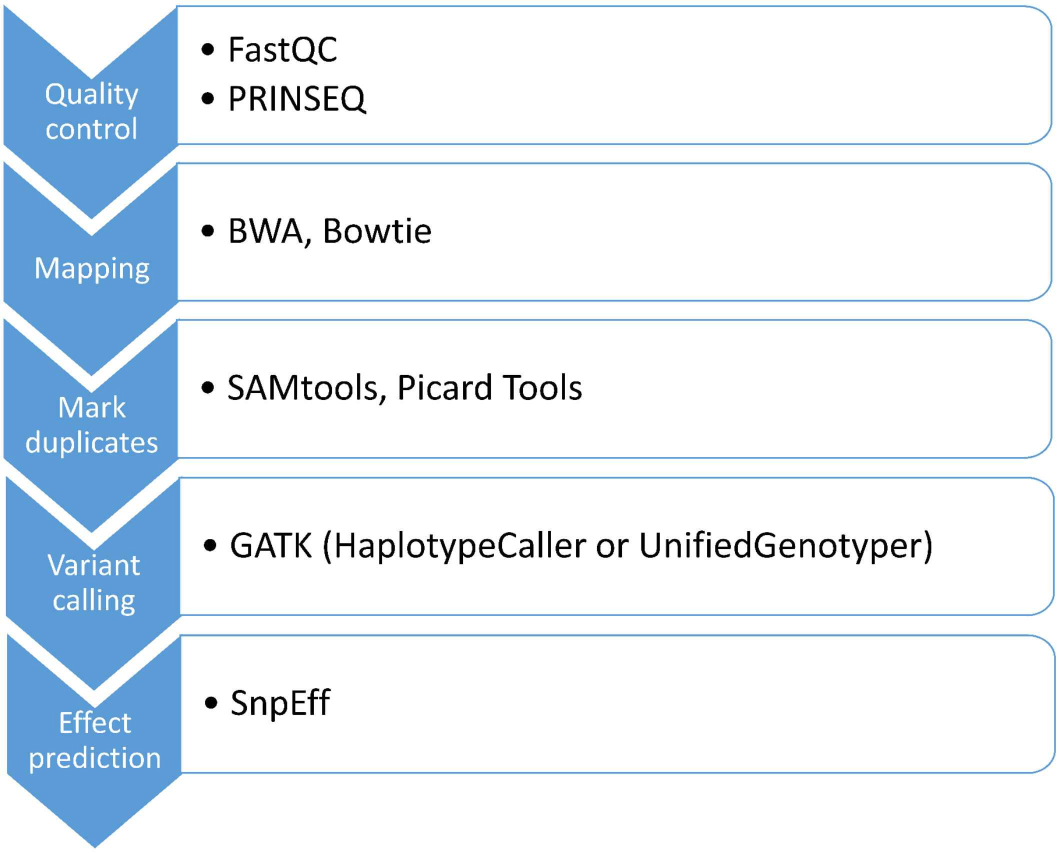

Figure 1. Common bioinformatics pipeline for variant calling in whole genome sequencing data generated by sequencing-by-synthesis technology.

Similar variant calling pipelines can be applied for RNA sequencing (RNA-seq) data (Conesa et al. 2016) (Fig. 2). However, in contrast to whole genome analyses, whole transcriptome analyses produce shorter sequence reads, are usually performed in specific tissues of interest and are therefore often focused on differential expression issues (Su et al. 2002). Pipelines for quantification of gene expressions usually use specific RNA-seq aligners for mapping of the short reads, followed by tagging of duplicates (Trapnell et al. 2009; Langmead & Salzberg 2012; Dobin et al. 2013). After that, reads are split into exon segments and trimmed for sequences overhanging into the intronic regions (McKenna et al. 2010). Finally, common pipelines apply RSEM tool for counting, which uses an expectation-maximization (EM) algorithm to estimate maximum likelihood expression levels (Li & Dewey 2011). The expected counts-matrix provided by RSEM, is further evaluated using tools like DESeq or edgeR for differential gene expression analysis (Robinson et al. 2010; Anders & Huber 2012).

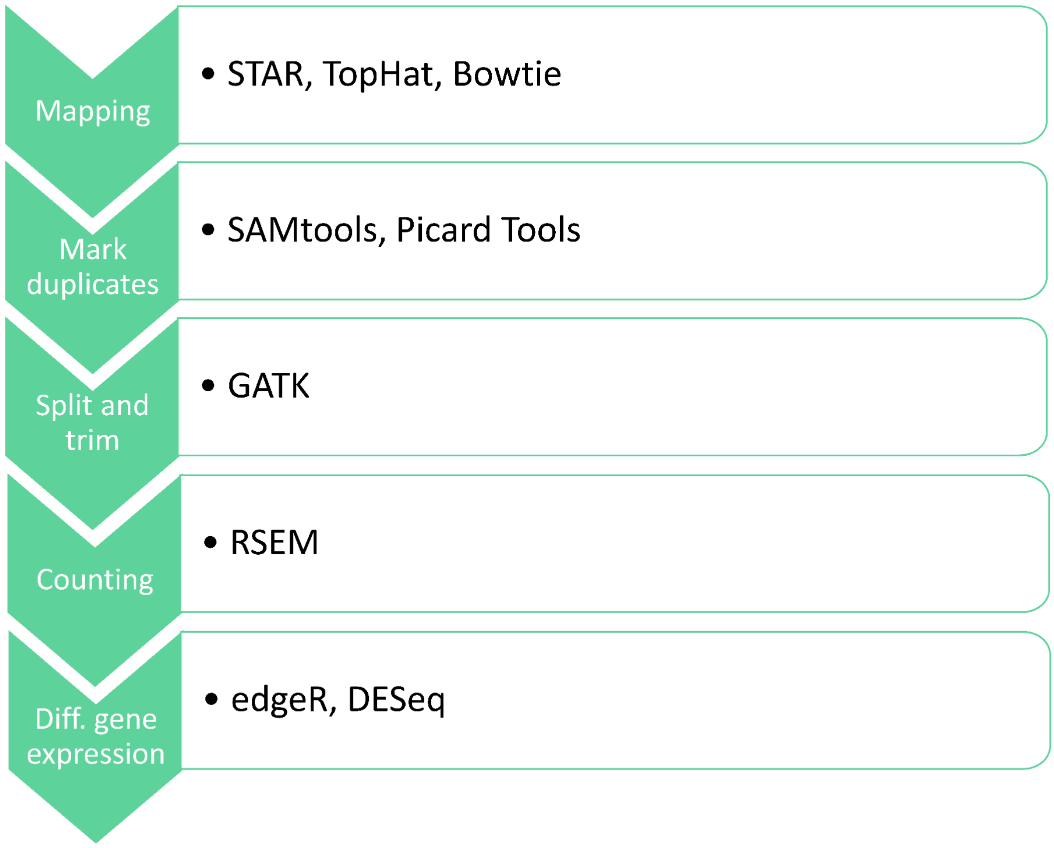

Figure 2. Common bioinformatics pipeline for quantification of gene expression using RNA sequencing data.

Despite these successful sequencing technologies of the next generation, the development of high-throughput sequencing has not stopped yet. A third generation of sequencing has evolved aiming at sequencing a single DNA molecule without fragmentation (Eid et al. 2009). This can

be accomplished by single-molecule real-time sequencing (PacBio, Pacific Biosciences, CA, USA) using four fluorescence-labeled dNTPs for synthesis and in parallel a real-time video detection of fluorescence signals (Eid et al. 2009). Another method is used by Nanopore technologies (Nanopore, Oxford, UK), which is based on current changes measured during the passing of a DNA molecule through a pore of one nanometer (Kasianowicz et al. 1996; Schneider & Dekker 2012). The key benefit of these single-molecule sequencers is the generation of long reads of about 10 kb (PacBio) and 50 kb (Nanopore) or more (Branton et al. 2008; Gordon et al. 2016; Jain et al. 2018). They allow the generation of large contigs and highly improve assembly statistics, in particular for de novo sequencing analyses (Frank et al. 2016). Nevertheless, single-molecule sequencing results in a comparatively low throughput and higher error rates (Lee et al. 2014). For this reason, most de novo sequencing projects are performed in a hybrid approach combining long reads from single-molecule sequencing with short and more accurate reads from sequencing-by-synthesis reactions (Branton et al. 2008; Lee et al. 2014). The substitution of 50% of short sequence reads with long read from PacBio was shown to result in a 3 fold increase of N50, an assembly metric designating the shortest contig at which half the total length of the assembly is made of contigs of that length or greater (Koren et al. 2012; Chaisson et al. 2015b).
Nevertheless, despite these advances of the third generation, NGS, especially the Illumina platform, has been proven to be the most cost-effective and accurate technology and is therefore still the method of choice especially for whole genome analyses based on known reference genomes in domestic animals (Reuter et al. 2015).

Domestic animal genomes

In the new era of genomics, the amount of genome data has been increased considerably. Various reference genomes of domestic animals have been built and made available in genome databases (NCBI: https://www.ncbi.nlm.nih.gov/genome; Ensembl: http://www.ensembl.org). They allow fast and accurate sequencing and evaluation of different breeds or populations of interest (Li et al. 2009b). Raw reads files derived from NGS are quality controlled and aligned to the reference sequence using structured programming frameworks (McKenna et al. 2010). That means, the quality and reliability of studies using whole genome sequencing data is highly dependent of the sequence coverage of reads at each individual base as well as reference accuracy itself (Harismendy et al. 2009). The house mouse assembly was found to be the most accurate one compared to the human reference assembly with notably low gap content (Seemann et al. 2015). In contrast, rather incomplete assemblies were identified in pig and dolphin with a high number of missing ultraconserved elements and protein-coding exons (Archibald et al. 2010; Seemann et al. 2015). Efforts have been made by different international communities to close those sequence gaps and errors in the reference assemblies to improve genome analyses. One of the latest achievements in updating of assemblies have been made for the cat reference genome Felis_catus_9.0 released in December 2017 as well as for the horse reference genome EquCab3.0 released in January 2018 (NCBI genome Annotation Release 103 Equus caballus, date of access: 13.02.2018). These upgrades of genomes display the increasing demand on genome reference-guided whole genome sequencing analyses for variant detection (English et al. 2012).

Table 1. Reference genomes in domestic animals. Current status of reference genome assemblies, their median length and number of annotated genes specified in NCBI database.

Common name (taxonomy, taxid ID)	Latest genome assembly	Median total length (Mb)	Number of protein-coding genes	Publications
Alpaca (*Vicugna pacos,* 30538)	Vicugna_pacos-2.0.2	2,092.95	19,510	Bactrian Camels Genome et al. (2012)
Arabian camel (*Camelus dromedaries*, 7461)	PRJNA234474_Ca_dromedarius_V1.0	2,055.06	19,410	Fitak et al. (2016)
Asiatic honeybee (*Apis cerana*, 7461)	ACSNU-2.0	228.332	10,608	-
Bactrian camel (*Camelus bactrianus*, 9837)	Ca_bactrianus_MBC_1.0	1,780.72	19,242	Bactrian Camels Genome et al. (2012)
Cat (*Felis catus*, 9685)	Felis_catus_9.0	2,841.07	19,748	O'Brien et al. (2002); Pontius et al. (2007)
Cattle (*Bos taurus*, 1758)	Bos_taurus_UMD_3.1.1	2,697.56	21,514	Bovine Genome et al. (2009); Zimin et al. (2009)
Chicken (*Gallus gallus*, 9031)	Gallus_gallus-5.0	1,125.64	19,137	International Chicken Genome Sequencing (2004)
Common canary (*Serinus canaria*, 9135)	SCA1	9,135	15,609	Frankl-Vilches et al. (2015)
Common carp (*Cyprinus carpio*, 7962)	Common carp genome	1,546.88	49,579	Ji et al. (2012)
Dog (*Canis lupus familiaris*, 9615)	CanFam3.1	2,254.63	20,039	Lindblad-Toh et al. (2005)

Table 1 continued.

Common name (taxonomy, taxid ID)	Latest genome assembly	Median total length (Mb)	Number of protein-coding genes	Publications
Domestic guinea pig (*Cavia porcellus*, 10141)	Cavpor3.0	2,723.22	19,976	-
Domestic silkworm (*Bombyx mori*, 7091)	ASM15162v1	397.687	13,798	Mita et al. (2004); Xia et al. (2004)
Donkey (*Equus asinus asinus*, 83772)	ASM130575v1	2,391.05	20,673	-
Ferret (*Mustela putorius furo*, 9669)	MusPutFur1.0	2,405.53	19,816	Peng et al. (2014)
Goat (*Capra hircus*, 9925)	Capra hircus ARS1	2,866.11	20,755	Bickhart et al. (2017)
Guppy (*Poecilia reticulate*, 8081)	Guppy_female_ 1.0+MT	731.622	22,897	Künstner et al. (2016)
Helmeted guineafowl (*Numida meleagris*, 8996)	NumMel1.0	1,043.26	16,101	-
Honeybee (*Apis mellifera*, 7460)	Amel_4.5	246.927	10,728	Weinstock et al. (2006)
Horse (*Equus caballus*, 9796)	EquCab2.0 (EquCab3.0)	2,426.22	21,129	Wade et al. (2009)
House mouse (*Mus musculus*, 10090)	Mus musculus GRCh38.p6	2,671.82	22,493	Mouse Genome Sequencing et al. (2002)
Pig (*Sus scrofa*, 9823)	Sscrofa11.1	2,457.91	20,790	Schook et al. (2005); Archibald et al. (2010)
Rabbit (*Oryctolagus cuniculus*, 9986)	OryCun2.0	2,737.46	20,547	Carneiro et al. (2014)
Rat (*Rattus norvegicus*, 10116)	Rattus norvegicus Rnor_6.0	2,743.3	23,347	Gibbs et al. (2004)
Rock pigeon (*Columba livia*, 8932)	Cliv_1.0	1,063.01	15,588	Shapiro et al. (2013)
Sheep (*Ovis aries*, 9940)	Oar_v4.0	2,738.01	20,645	International Sheep Genomics et al. (2010)
Swan goose (*Anser cygnoides*, 8845)	Anser cygnoides domesticus	1,124.72	20,787	-

Table 1 continued.

Common name (taxonomy, taxid ID)	Latest genome assembly	Median total length (Mb)	Number of protein-coding genes	Publications
Turkey (*Meleagris gallopavo*, 9103	Turkey_5.0	1,128.34	18,511	Dalloul et al. (2010)
Water buffalo (*Bubalus bubalis*, 391902)	UMD_CASPUR_WB_2.0	2,921.06	21,711	-
White-rumped munia (*Lonchura striata*, 40157)	LonStrDom1	1,060.17	17,446	Colquitt et al. (2018)
Zebu cattle (*Bos indicus*, 9915)	Bos_indicus_1.0	2,673.97	19,900	Canavez et al. (2012)

Detection of SNPs and SNVs

NGS has opened up the way for fast detection of common single-nucleotide polymorphisms (SNPs) as well as novel single-nucleotide variants (SNVs) distributed all over the genome. Millions of SNPs/SNVs were identified in domestic animals ranging from 2-3 M in single individuals (Eck et al. 2009; Doan et al. 2012) to about 27 M in larger sample sizes used in the 234 bull genomes sequencing project, in which 78.4% of the SNPs/SNVs were classified as novel (Daetwyler et al. 2014).

A small fraction of these variants is located in protein-coding regions. In particular non-synonymous single nucleotide polymorphisms (nsSNPs), a type of genetic mutations in protein-coding regions that cause a single amino acid substitution in a protein sequence, have been of special interest in investigations of domestic animals, as they are likely to affect protein function and might have an impact on economically important phenotypic traits (Kim et al. 2003; Kumar et al. 2009). About 26-32% of human nsSNPs in clinically important genes were predicted to have functional consequences (Chasman & Adams 2001). It was postulated that nsSNPs occur at a significantly lower rate and with a lower allele frequency than synonymous substitutions as a result of selection acting against deleterious alleles during evolution (Cargill et al. 1999). This low frequency of nsSNPs was also observed in other mammals like cattle, in which 87% of amino-acid changing mutations were estimated to be deleterious (MacEachern et al. 2006). The prediction of these functional effects of nsSNPs on the protein in every organism is based on the level of conservation, which is assessed by various effect prediction tools (Kumar et al.

2009; Bendl et al. 2014). The higher the conservation of specific SNP-positions is, the less a substitution of an amino acid is potentially tolerated (Kumar et al. 2009). In contrast, synonymous SNPs (sSNPs) that do not cause a single amino acid substitution, are usually predicted as tolerated variants without any effects on protein function (Ng & Henikoff 2003; Kimchi-Sarfaty et al. 2007). Nevertheless, the increased detection rates of SNPs/SNVs by NGS data analyses have more and more revealed functional effects of sSNPs on protein folding, especially co-translational folding, or mRNA stability (Cortazzo et al. 2002; Capon et al. 2004; Kimchi-Sarfaty et al. 2007). It was shown that nsSNPs and sSNPs shared similar likelihood and effect size for disease association and should thus be both considered in functional studies (Chen et al. 2010).

However, despite this significant importance of coding-region SNPs, the number of non-coding variants with functional effects is as large as nsSNPs and sSNPs (Consortium 2012). Human genome data showed that only 2.94% of the genome is covered with exons of protein-coding genes. It was estimated that 90% of trait-associated SNPs are located in non-coding regions (Chen et al. 2016). These SNPs usually affect functional elements, including regions of transcription factor association, promoter regions, enhancers, silencers, DNA-methylation regions or histone modifications (Narlikar & Ovcharenko 2009; Consortium 2012). All these types of regulatory elements were systematically assigned in the Encyclopedia of DNA Elements (ENCODE) project for human genomes, which served as a pioneer for discovering all functional elements of the genome (Consortium 2012). It revealed 80.4% of the genome to be involved in biochemical RNA- and/or chromatin-associated events. NGS technology enabled massively parallel sequencing of ChIP (chromatin immunoprecipitation) libraries in order to identify regions bound by proteins (Hu et al. 2010). In particular, transcription factor binding sites were identified mostly containing strong DNA-binding motifs, although the sequences of the motifs were highly species-specific (Schmidt et al. 2010; Consortium 2012). It was proposed that various species show transcription factor bound regions near genes, whose expression levels are dependent of transcription factors, but do not necessarily have the same motifs (Schmidt et al. 2010). SNPs/SNVs modifying the accessibility for these regulatory factors by affecting allelic chromatin states, were found to be concentrated in DNase I hypersensitivity regions (Neph et al. 2012; Light et al. 2014). The loss of the condensed chromatin structure in these regions resulted in an enrichment of DNase I cleavage but a punctuated protection from cleavage in regions bound by regulatory factors designated as 'footprints' (Neph et al. 2012). Heterozygous SNVs were supposed to confer significant allelic imbalance in chromatin accessibility by disrupting the footprinting of DNA-binding motifs.

Thus, the expansive mapping of regulatory elements in the ENCODE project by genome sequencing technology elucidated functional sequence compartments as potential targets for SNVs or SNPs. In domestic animals, efforts have been made to follow the example of ENCODE by the promotion of the Functional Annotation of ANimal Genomes (FAANG) project, an international project for assembling comprehensive maps of functional elements in the genomes of domesticated animals (Andersson et al. 2015). Due to the complexity of functional elements and the inter-species differences, it was outlined as an invaluable resource for the identification of genotype-phenotype relations (Barbosa-Morais et al. 2012; Yue et al. 2014; Andersson et al. 2015). In 2016, the FAANG Consortium established reference datasets under pilot projects, initiated plans for sample collection and specified data analysis standards (Tuggle et al. 2016). This project shall contribute to understanding functional effects of variants, especially of those located in non-coding regions.

However, despite these efforts to specifically identify SNVs/SNPs and their functional effects, these type of variants are also considered as useful tools for genetic analyses (Lee et al. 2006). Various studies used SNPs as genetic markers equally distributed over the genomes for mapping of desired traits (Petersen et al. 2013; Daetwyler et al. 2014). In particular, genome-wide association studies (GWAS) and analyses for selection signatures were applied for SNP-genotyping data to identify candidate regions (Burren et al. 2016; Gottschalk et al. 2016). In sliding window approaches runs of homozygosity (ROH) detection was performed in order to identify long stretches of consecutive homozygous SNP genotypes as potential selection signals (Szpiech et al. 2013). These analyses were run on basis of bead chip genotyping results and further validated using whole genome sequencing data or directly run on SNVs/SNPs derived from NGS data (MacLeod et al. 2013; Reinartz et al. 2017). The identified ROHs were directly explored for causal variants found in genome data and were also widely used for population genetic approaches in order to characterize breeds or specific populations (Bosse et al. 2012; Purfield et al. 2012; Wang et al. 2015). To investigate the reduced genetic variability in homozygosity regions for indications of positive selection or inbreeding, group comparative population studies additionally performed analyses for selective sweeps, a reduction of variation among nucleotides surrounding a mutation (Pfahler & Distl 2015). These regions provided evidence for differences in selection and for particular specialization of breeds (Rubin et al. 2010).

The more the genome was covered with SNPs, the higher was the accuracy of detection results (Cai et al. 2011). To avoid cost-intensive NGS of large sample sizes, low density genotyping data were in some studies imputed on SNPs derived from whole genome sequencing analyses and used to further narrow down regions of interest (van Binsbergen et al. 2014). The specificity of these imputed sequence variants was estimated to be high in total, but showed limitations for SNPs with low minor allele frequencies (MAF) or poorly covered genome regions (Daetwyler et al. 2014).

This demonstrates the versatile use for SNVs/SNPs from whole genome sequencing data not only for the identification of candidate variants for individual traits but also for targeting regions of interest and signatures of potential selection for domestic animal breeds.

Insertions/deletions and copy number variations

Since the introduction of high throughput sequencing technologies, various studies focused on SNPs as the main genetic basis for variation among mammals. Nevertheless, more and more analyses of whole genome sequencing data revealed a high frequency of insertions/deletions (indels) as well as structural variation in the genome and emphasized their contribution to diversity and disease traits (Weber et al. 2002; Feuk et al. 2006; Chen et al. 2009). Indels are defined as variations of 1 bp to 1 kb, whereas structural variation comprise deletions, duplications and large-scale copy-number variants (CNVs) >1 kb (Weber et al. 2002; Feuk et al. 2006).

In domestic animals, various research groups identified indels or structural variation affecting breed specific phenotypes or performance traits (Bickhart et al. 2012; Axelsson et al. 2013). However, it was shown that a comprehensive assessment of particularly larger deletions, insertions or translocations was more challenging due to their complexity and limitations of mapping of short sequence reads (Ye et al. 2009; Alkan et al. 2011). A thorough characterization of structural variations was proposed to be fully accomplished by de novo assembly, which was found to be more efficient by using long sequence reads derived from single-molecule sequencing technology (Schadt et al. 2010; Chaisson et al. 2015a). These long reads enabled to close so far unsolved gaps and the detection of larger, complex regions in the genome (Chaisson et al. 2015a).

Despite these advantages of third generation sequencing technologies, the production of long sequence reads has been found to be cost-intensive and thus not applicable for larger sample sizes in particular in animal studies. That is why efforts have been made to improve structural variant detection, in particular the identification of CNVs. CNV analyses have considerably made progress since the implementation of genome analyses, initially based on array comparative genomic hybridization (CGH) methodology (Pinkel et al. 1998), followed by SNP genotyping chips (Winchester et al. 2009) and CNV detection in whole genome sequencing data (Wheeler et al. 2008). Basically, four different methods have been used for NGS data to evaluate CNVs, which all have their advantages and limitations (Pirooznia et al. 2015). One of the first approaches based on NGS data has been the read-pair method, an analysis of read-pair distances in the investigated data compared to expected distances according to the reference sequence (Korbel et al. 2007). It was shown that the average breakpoint resolution was high, although some unmated insertions, lacking one predicted breakpoint, were detected, whose sizes could not be determined (Korbel et al. 2007). The high information content derived from paired-end sequencing, a sequencing of both ends of a fragment, was not only used for read-pair but also for split-read approaches. This method allows pinpointing the exact breakpoints of structural variation, by using the information of unmapped or partially mapped reads whose mate was reliably mapped, so-called hanging read pairs (Zhang et al. 2011; Jiang et al. 2012). Thus, the actual sequence content of an insertion can be detected by this approach, whereas long structural variations are more difficult to detect (Ye et al. 2009; Zhang et al. 2011). However, it was shown that these variants, which were refractory to paired-end based tools, were captured by the read-depth method (Yoon et al. 2009). This type of analysis based on the read depth at each position in the genome in relation to a control sample or a bunch of samples, enables the detection of segmental duplications, as it is not dependent of uniquely mapping reads in a region to estimate the coverage (Yoon et al. 2009). An also independent method, the assembly method, produces contigs of assembled whole genome sequencing data in a first step and compares these results to a reference genome (Nijkamp et al. 2012) . After all, the most effective way to identify CNVs is to combine different detection methods in order to increase detection accuracy (Rausch et al. 2012). Various tools apply joint analyses using paired-end insert size detection, identification of discordantly mapped read-pairs, chimeric reads and read depth estimation in two- or three-step runs (Medvedev et al. 2009; Ye et al. 2009; Alkan et al. 2011; DePristo et al. 2011; Rausch et al. 2012; Narzisi et al. 2014; Rimmer et al. 2014; Li 2015; Chong et al. 2017; Yang et al. 2018). The continuing development of tools for detection of structural variations and the increasing knowledge of its relevance for disease and performance traits has more and more supported the standard integration of structural variant detection in genome analysis of domestic animals.Thus, various studies on different species demonstrated the essential role of SNPs/SNVs as well as indels and structural variations. Their significance for complex interactions in the genome and their role in targeted selection of domestic animals still need to be further investigated.

5 Concepts and Objectives

This habilitation thesis focusses on the area of genomics applied in domestic animals. The individual studies aim at different aspects of this field, investigating structure, function and evolution of genomes using NGS technology. The key link in-between all these analyses is the strong-targeted human selection of animal populations or breeds in the course of domestication that are traced back on the domestic animal genomes.

The following objectives were on focus in this work:

1. Characterizing genomic variations with regard to diversity in highly selected breeds and less selected populations.
2. Identifying runs of homozygosity as potential signatures of selection playing an essential role in the evolutionary processes.
3. Detection of variants in highly selected regions on the genome that play a role in disease development.
4. Identifying functional variants and synergistic effects on the genome that affect characteristic breed specific phenotypic traits.

6 Results and Discussion

6.1 Characterisation of horse genomes under strong targeted selection using Next Generation Sequencing

The fundament of genome analyses in domestic animals is built on high quality reference genomes as well as the knowledge of known or novel variants and their functional effects (Doan et al. 2012; Daetwyler et al. 2014). In the presented **publications 1 and 2** whole genome sequencing data from different types of horses were investigated in order to asses the distribution of variants, their detection accuracy given a certain genome coverage and their involvement in selection signatures. At the time of first investigations in **publication 1**, only two studies on NGS data in horses were published, dealing with the genome of a Quarter Horse mare as well as with a group of seven equids including five domestic horses, one Przewalski and one donkey (Doan et al. 2012; Orlando et al. 2013). It was still unclear how the detected variants in these horses were actually distributed among horse populations and how selective pressures were involved in the occurrence of known or novel variants. Thus, we analyzed NGS data from horses breeds highly selected for comformation or performance traits and compared them to data from non-breed horses, which were less exposed to human influences, in order to approach these questions.

A coverage of 10.97X-25.38X (**publication 1**) and 5.92X-35.18X (**publication 2**) of the individual samples allowed an accurate detection of known variants affecting coat color, body size or performance traits. Based on bead chip genotyping data as reference, the accuracy of variant calling results was 95.92% in the Arabian horse and 97.42% in the Hanoverian. Further investigation of four Hanoverian and one Arabian in **publication 2**, revealed in total 22%-26% of the heterozygous SNPs to be false-negatively detected as homozygous in NGS data. Only few SNPs (19-49 SNPs) were estimated to be false-positive and thus heterozygous in NGS data, although designated as homozygous in bead chip analysis. These sequence error estimations confirmed a reliable detection of variants in this coverage range. As it was proposed for analyzed whole-genome sequencing data of a Quarter Horse mare, the false-negative rate is comparatively high, if variants are stringently filtered for their individual read depth (Doan et al. 2012). For this reason, we used raw variant calling results for the estimation of potential novel variants. Approximately 5,000,000 SNPs and 900,000 indels were identified in **publication 1** in the individual horses including two Hanoverian, one Arabian, one Duelmen Horse and one Sorraia. We found an overlap of 907,776 SNPs in comparison to dbSNP variant database records (NCBI) and an overlap of 907,864 SNPs with Ensembl variant database. A similar number of SNPs (910,822) was identified overlapping with variants uploaded in the database from Broad Institute (horse_snp_release/v2). In total, 3,394,883 SNPs were identified neither in public databases nor in previous studies on whole-genome sequencing data in horses and were for this reason designated as novel variants. Furthermore, we found 868,525 novel indels. These high numbers of novel SNVs and indels showed that millions of variants have been undetected so far. It provides evidence of the high genetic variability in-between different horse populations or breeds as it has already been supposed in bead chip based analyses (Petersen et al. 2013).

This suggestion was supported in **publication 1** by the analysis of private variants for their functional effects in breed horses (Hanoverian and Arabian) as well as non-breed horses (Duelmen Horse and Sorraia). Variants private for non-breeds were enriched with genes involved in metabolic pathways and morphogenesis. We assume that the survival of these horses, which are kept under free-range conditions, is highly dependent on fast development, efficient digestion and energy intake as it was shown in the Przewalski horse (Kuntz et al. 2006). In contrast, modern horse breeds are specifically selected for performance or conformation traits (Koenen et al. 2004). We found indications for this high specialization of breed horses in coding regions with private variants. They were enriched with genes affecting muscle contraction, neurological system processes, ectoderm development and ion/cation transport. In addition, genes involved in the development of the immune system were more frequent in private indels with codon changing effects in breed horses in comparison to non-breed horses.

In order to characterize these potential signatures of selection in breeds or non-breeds, we performed runs of homozygosity (ROHs) analyses on these whole-genome data. The detection of ROHs was shown to be a powerful approach for the estimation of inbreeding and the population's demographic history (Curik et al. 2014). Especially genome signatures of artificial selection were successfully identified in groups of individuals specifically bred for performance or conformation traits (Kim et al. 2013). Based on the high number of SNPs identified in the ten investigated horses in **publication 2**, we detected an average number of 3492 ROHs in sliding windows of 50 homozygous SNPs. Further approach for a minimum of 500 homozygous SNPs revealed an average number of 292 ROHs. The genomic inbreeding coefficient (FROH) was found to be significantly high in the Thoroughbred and in the Sorraia horses. This supported previous suggestion of a reduced effective population size in the endangered Sorraia horse and an increase of inbreeding in the Thoroughbred over the past years (Luis et al. 2007; Binns et al. 2012).

Further investigation of the Hanoverian breed for potential selection signatures revealed 18 ROHs shared by all four Hanoverian, which were not identified in any other of the investigated horses. As expected from the high selection for performance of these sports horses, we found genes in ROH regions, which were proposed to play a role in osteogenic proliferation and differentiation (*ALX homeobox 4*), neurologic control (*cysteine sulfinic acid decarboxylase, dyslexia susceptibility 1*) and signal transduction (*TBC1 domain family, member 30*) (Wu 1982; Antonopoulou et al. 2004; Ishibashi et al. 2009; Darki et al. 2012). Furthermore, the whole group of investigated Hanoverian, Arabian, Saxon-Thuringian Heavy Warmblood and Thoroughbred revealed a private ROH in the region of *V-Yes-1 Yamaguchi Sarcoma Viral Oncogene Homolog 1 (YES1)*, whose expression was shown to be activated in response to testicular heat stress in mice and thus suggested to be important for pachytene spermatocytes in the prophase of meiosis (Liang et al. 2013). This evidence for positive selection for reproduction was not only found in breed horses but also in non-breeds. The *KIT ligand (KITLG)* gene, which is known to play an important role in male fertility and spermatogenesis, was detected in the region of three consensus ROHs identified in all ten horses, breeds and non-breeds (Blume-Jensen et al. 2000; Runyan et al. 2006). These findings support the assumption that functionally important regions under natural or artificial selection, can be traced back by their specific selection signatures and are relevant for highly selected horse breeds as well as non-breeds (Qanbari et al. 2011).

However, the use of artificial insemination applied widely in horse breeding, has in particular strongly influenced positive reproductive factors in stallions in breed horses (Hamann et al. 2007). In the presented **publication 3**, we screened whole genome sequencing data from seven stallions and four mares of different breeds for high-impact variants involved in stallion fertility. We assumed that variants affecting genes involved in male reproductive processes were less likely to show a homozygous mutant genotype due to strong selective pressures for high fertility. This was supported by investigations of other species, which gave evidence for a decrease of the reproductive efficiency or even infertility due to loss-of-function mutations (Sonstegard et al. 2013; Pausch et al. 2014; Fujiwara et al. 2015). Our filtering analysis for variants, which showed no homozygous mutant genotype in all investigated horses and were located within one of 1194 identified equine genes orthologue to human or mice fertility related genes, resulted in 1259 variants in coding regions. Further screening of these data for variants with putative high impacts according to SnpEff predictions (Cingolani et al. 2012) resulted in 19 SNPs. In total 16 of these SNPs were located in 15 genes involved in reproductive processes, whereas further three genes were designated as novel or pseudogenes with unknown function

(Ensembl database). Validation of these 19 SNPs in 96 stallions revealed two variants to be monomorphic, which were thus not further investigated for this reason. The remaining 17 SNPs were genotyped in 337 horses including 226 Hanoverian stallions from the National State Stud Celle and 111 stallions from different breeds, whose fertility was ensured by records of a minimum of 10 viable offspring. Nine SNPs located in the reproduction-related genes cystic fibrosis transmembrane conductance regulator (*CFTR), oviductal glycoprotein 1 (OVGP1), F-box protein 43 (FBXO43), testis specific serine kinase 6 (TSSK6), polycystic kidney disease-associated protein (PKD1), forkhead box P1 (FOXP1), T-complex 11 (TCP11), SPATA31 subfamily E member 1 (SPATA31E1)* and *neurogenic locus notch homolog protein 1 (NOTCH1)* did not show a homozygous mutant genotype in the investigated stallions. Additional GWAS for all 15 SNPs located in male reproduction (novel genes excluded) related genes in 216 Hanoverian stallions with defined breeding values of the paternal component of the pregnancy rate per estrus (EBV-PAT), revealed a significant association with stallion fertility in the *NOTCH1* variant (g.37453246G>C). These results indicated that *NOTCH1* high-impact variant is a potential stallion fertility-associated locus in Hanoverian stallions. It was postulated that *NOTCH1* is required for spermatogenesis in rat and human testes and thus is crucial for differentiation or survival of male germ cells (Azuma et al. 2002). In addition, we found *NOTCH1* interacting with the identified genes harboring potential high-impact variants with absent homozygous mutant genotypes, suggesting these 9 mutations also as deleterious factors for stallion fertility. This supports the assumption that homozygous mutant genotypes promote infertility in stallions, which is underlies negative selection in the breeding populations. NGS analysis in **publication 3** allowed the identification of these potential high-impact variants by screening whole genomes and thus demonstrated that also rare variants could be accurately detected.

All three studies in **publication 1, 2** and **3** highlight the detection of genomic changes because of targeted specific selection of horse breeds. This targeted selection has resulted in particularly diverse breed types, which vary widely not only by their performance but also by their external appearance. One of the most apparent characteristics of horse breeds is their body size, ranging from a height at the withers (transition from neck to back) of 74 cm (29.13 inches) or smaller to a height of 183 cm (72 inches) or higher (Petersen et al. 2013). Even within the group of ponies, there is a wide range of withers height from miniature size (smaller than 87 cm/34.25 inches) up to a maximum height of 148 cm (58.27 inches). In the presented **publication 4**, we investigated potential signatures of selection for extremely small body size in miniature horses. For this purpose, we analyzed whole-genome sequencing data of two Miniature Shetland ponies and one standard-sized Shetland pony for signatures of selection involved in size development. The Shetland pony represents an optimal model for the investigation of selection signals, because this small pony breed was further subdivided into Miniature Shetland ponies, which are not allowed to exceed 87 cm (34.25 inches) height at the withers, and standard-sized Shetland ponies with a withers height up to 107 cm (42.13 inches), which can only differentiated by their size (Molra C. Reeve 2011). We assume that selection for body size in Shetland ponies was targeted at both directions, specifically developing miniature and standard-sized types. For this reason, we performed detection of shared ROH regions found in Miniature and standard-sized Shetland ponies. We identified 460 ROH regions spanning in total 1492 genes common in all three individuals. Further functional investigations revealed one gene, *high mobility group AT-hook 2 (HMGA2),* in a shared ROH region, which was suggested to harbor a variant in a gap region of the reference genome, involved in body size development in ponies (Frischknecht et al. 2015). We confirmed this *HMGA2* variant (c.83G>A) in the Shetland pony. However, it could not fully explain the size differences in-between miniature and standard size.

Thus, all shared ROH regions were screened for variants with predicted high or moderate effects according to SnpEff predictions (Cingolani et al. 2012) and revealed 1364 variants of which 216 had a potential deleterious effect. Nine of the 216 variants were classified as growth, muscle development or skeletal development related genes by DAVID functional annotation tool (Huang da et al. 2009a, b). One of these variants, located in *disintegrin-like and metalloprotease with thrombospondin type 1 motif 17 (ADAMTS17),* and a second missense variant in *osteocrin (OSTN),* showed a homozygous mutant genotype in the three Shetland ponies. Furthermore, the mutant allele of these two variants could not be identified in whole-genome sequencing data from 29 control equids, including modern horse breeds, non-breeds, Przewalski horses, one donkey and also ancient Scythian horses from Berel' (Orlando et al. 2013; Librado et al. 2017). Validation of all ten identified SNVs in Shetland ponies of different heights and further samples of various horse breeds revealed three missense variants located in *growth hormone 1 (GH1), ADAMTS17* and *OSTN* to be highly associated with withers height. Moreover, the joint genotypes of these three variants and the additionally validated *HMGA2* variant could explain the height differences of miniature and standard-sized Shetland ponies and Shetland pony-related breeds. We conclude that the synergistic effects of these four variants are responsible for the determination of miniature size, a height at the withers of smaller or equal to 87 cm (34.25 inches), in horses, explaining 72% of the size variation. As the variants were specific for Shetland ponies and related breeds, it confirmed the suggestion that they occurred in the Shetland pony breed and resulted in a miniature type due to strong-targeted selection.

Based on the size of the homozygosity regions, we found indication that the first mutant allele arose over 1000 years ago. Samples from ancient genomes of Scythian horses from the Iron Age (Librado et al. 2017), investigated in our study, did not harbor any of these variants. This leads to the suggestion that miniature size in horses is a product of more recent human selection, presumably due to the advantages of small size for Shetland ponies as workhorses in the coalmines or for optical purposes (Van Bergen & Van Arendonk 1993; Molra C. Reeve 2011). However, the pony-type, ranging from >87 cm (34.25 inches) to 148 cm (58.27 inches) was presumably already present in early domestic horses. It was shown that whole-genome sequences from Iron Age stallions from the earliest Scythian royal mound of Arzhan I as well as Scythian stallions from the kurgan 11 of Berel' did all harbor the pony-associated genotype of the *LCORL* variant, which was described as the main regulator for body size differences in-between ponies and larger horses (Makvandi-Nejad et al. 2012; Librado et al. 2017). In the presented **publication 4**, we found evidences that supported this assumption.

We performed copy number variant (CNVs) detection, in order to identify Shetland pony specific variations. In total 91 CNV regions harboring losses and 6 CNV regions harboring gains were found in whole-genome sequencing data of the three Shetland ponies in comparison to three Hanoverian, two Arabian, one Thoroughbred, three non-breed horses, one Marwari horse, one Standardbred and one Saxon-Thuringian Heavy Warmblood. Validations of a 7245bp deletion spanning *diaphanous related formin 3 (DIAPH3)* in various samples of equids revealed this loss to be exclusive in Shetland ponies, Shetland pony-related breeds, Icelandic horses, Lewitzer, Welsh horses and donkeys. Furthermore, this deletion was also confirmed in a Scythian horse. We assume that this CNV represents a pony-specific variant, potentially involved in traits characterisitc for the pony breeds.

The findings once more underline the role of body size as an interesting model of rapid evolutionary changes during horse domestication (Brooks et al. 2010b). Whole-genome sequencing data allowed a deep insight into genomic changes, which shaped the horse populations due to the strong influence of targeted selection.

All four **publications 1, 2, 3** and **4** emphasized the high diversity of horse populations displayed by the extremely high number of private SNPs, indels and CNVs and thus highlighted the value of NGS data for versatile characterization of genomes.

6.2 Analysis of dog genomes for breed specific selection signatures associated with disease traits

The dog represents an extremely divers species, represented by various breeds highly selected for their appearance, abilities and behavior (Akey et al. 2010). Intense selection for sharply defined breeding objectives and targeted inbreeding has supported the development of breed specific variants in domestic dog genomes, promoting not only desired phenotypes but also disease susceptibility traits (Shearin & Ostrander 2010).

In the presented **publications 5, 6** and **7** (appendix 12.4, 12.5 and 12.6) we performed ROH analyses for potential selection signatures as means of identifying causative variants for breed-specific diseases. This approach was used to respond to the main challenge for NGS-based analyses: Huge datasets provide huge numbers of variants, of which only few can be considered as causative for the investigated diseases (Amos et al. 2011). For targeted filtering of candidate variants, genomic regions are narrowed down by the identification of potential selection signals, often further accompanied by GWAS. In **publication 5**, we performed both ROH analysis and GWAS in Norwegian Lundehund dogs in order to identify a recessive major gene for the Lundehund syndrome (LS). This severe syndrome is a complex disease composed of protein-losing enteropathy (PLE), intestinal lymphangiectasia, gastrointestinal disturbance, inflammatory bowel disease and malabsorption (FlesjÅ & Yri 1977; Kolbjørnsen et al. 1994). GWAS based on bead chip genotyping data revealed a significant association on canis familiaris autosome (CFA) 34 in the region of 23,373,982-24,488,983 bp (CanFam 3.1). Subsequent whole-genome sequencing of pools of three LS-affected and three LS-unaffected dogs resulted in more than 10 million SNPs/SNVs and more than 8 million indels. After stringent quality control, 8,815,395 SNPs/SNVs (LS-affected) as well as 8,799,793 SNPs/SNVs (LS-unaffected) remained for ROH analysis. To narrow down the number of potential candidate variants for LS, we performed ROH detection for common ROH regions in LS-affected dogs. In total, three of the 1713 identified ROHs were found at 23,223,356-24,490,050 bp, overlapping with the region of genome-wide association. In addition, we detected 600 consensus ROH regions, which showed homozygous stretches in both pools comprising LS-affected and LS-unaffected Lundehund. One of these ROH regions harbored the gene *nucleotide-binding oligomerization domain containing 1 (NOD1)*, which was shown to play a role in inflammatory bowel disease in human (McGovern et al. 2005) and thus suggested an involvement in breed-disposition for LS. However, no LS-specific missense variant, splice variant or other variants with high or moderate effects according to SnpEff predictions (Cingolani et al. 2012) were identified in consensus or in private ROH regions. Further filtering performed for all variants derived from whole-genome sequencing analysis, revealed 32 SNPs and one indel, which were homozygous mutant in LS-affected dogs and heterozygous or homozygous wild-type in LS-unaffected dogs as well as in five control sequences from a Korean Jindo Dog, an Afghan Hound, a Border Collie and two German Shepherd dogs. One of these variants was located 1.2 Mb proximal of the region of GWA in the gene *LEPREL1*. Validation of this missense variant in 36 Lundehund and 186 dogs of different breeds confirmed a significant association with LS and revealed all Lundehund with diagnosed LS to harbor the homozygous mutant genotype for this variant. The results of this study in **publication 5** show that the whole genome sequencing analysis accompanied by GWAS and ROH detection allowed a successful identification of *LEPREL1* variant as causative variant for LS.

A similar approach was chosen in **publication 6** and **7** for a breed-specific disease in Shar-Pei dogs. The Shar-Pei Autoinflammatory Disease (SPAID) is assigned as a syndrome with typical signs of inflammation, including swollen hooks, dermatitis, otitis and fever (Olsson et al. 2013). Due to its high similarity to human Mediterranean fever, the Familial Shar-Pei Fever was initially suspected to be a separate disease potentially triggered by a CNV located upstream of *hyaluronic acid synthase 2 (HAS2)* (Olsson et al. 2011). In **publication 6**, we investigated the distribution of this CNV (CNV-1), which was proposed to be causative for Familial Shar-Pei Fever, and a second overlapping CNV (CNV-2) in this region. The detection results of copy numbers in 95 Shar-Pei dogs and 14 controls refuted an association of CNV-1 with SPAID. In addition, we investigated the hypothesis of an association of CNV-1 with the development of a strongly wrinkled skin and a heavily padded muzzle, characteristic for the meatmouth Shar-Pei type (Olsson et al. 2011). This assumption was also not confirmed in our analysis. Nevertheless, we found the development of wrinkles closely related to Familial Shar-Pei Fever.

After further clinical studies of Swedish Shar-Pei dogs, it became clear that the Shar-Pei Fever was one of many more signs of a complex syndrome, the SPAID (Olsson et al. 2013). In the presented **publication 7**, we performed whole genome sequencing in two SPAID-affected Shar-Pei in order to identify the causal variant for this syndrome. In total 27 variants were detected, which were exclusively homozygous in the two SPAID-affected Shar-Pei but homozygous wild-type in five dogs from different breeds, which have already been used as controls in **publication 5**. Validation of six potentially deleterious SNVs in 102 SPAID-affected and 62 SPAID-unaffected Shar-Pei revealed one variant with a borderline significant P-value in *transforming growth factor beta receptor 3 (TGFBR3)* and one highly significant variant in *MDM2 binding protein (MTBP). MTBP* was found to be located proximal of the previously investigated CNV-1. Furthermore it was close by the significant peaks of GWAS, which was performed for the individual signs fever, arthritis, vesicular hyaluronosis, otitis and amyloidosis in Shar-Pei and revealed association signals in the region of 23,180,227-27,371,905 bp on CFA13.

Due to the late onset of the disease, we carefully investigated those Shar-Pei dogs for *MTBP* variant *MTBP*:g.19383758G>A that were older than six years. It was found to be homozygous mutant (A/A) in SPAID-affected dogs, whereas SPAID-unaffected dogs harbored the wild-type genotype (G/G). In the total group of 25 SPAID-affected dogs with a heterozygous genotype (G/A), 17 showed clinical signs of SPAID, whereas 8 individuals were not affected at the time of investigation. We proposed that this distribution supported an incomplete dominant mode of inheritance for SPAID. The *MTBP* variant was suggested to affect the binding capacity of MTBP resulting in an increased binding to MDM2 or a reduced dissociation of this complex (Alam et al. 2012). Our expression analysis of cellular tumor antigen p53 (Tp53) as bio-indicator for MTBP-MDM2-activity supported this assumption and revealed a decreased expression of TP53 due to an increase MTBP-MDM2 activity in SPAID-affected dogs. We assumed that this complex supports the pro-inflammatory effect of MDM2 similar as it was observed in human Mediterranean Fever (Mulay et al. 2012; Manukyan et al. 2015).

Additional ROH analysis based on bead chip data confirmed a potential selection signature in this region. The largest ROH region was located on CFA13 at 18,625,926-19,967,354 bp spanning *MTBP*. Due to the frequent occurrence of SPAID in Shar-Pei with a meatmouth and a heavily wrinkled skin, we assumed that the causative variant for both traits were located close by and thus inherited together. Strong targeted selection for the desired meatmouth type might have promoted thereby disease development.

The results of both studies in **publication 5, 6** and **7** show that intra-breed homogeneity supports genome analyses for specific traits by the use of ROH detection and GWAS. Selection signatures for desired phenotypes can be concurrently associated with disease traits and therefore occur frequently in the breed due to strong human selection for the intended breeding goal.

6.3 Investigation of variants for specific phenotypes in whole genome sequencing data of domestic animals

Whole genome sequencing provides a multitude of variant data, comprising mutations with high, moderate, low or no functional effects (Cingolani et al. 2012). The detection results of NGS data can be scanned for potential candidate mutations associated with desired phenotypes or disease traits, superseding the previous single candidate gene-sequencing approaches (Zuryn et al. 2010). This allows fast and accurate investigation of variants in various trait related genes, neighboring non-coding regions and in further genes, which have not been associated with this trait so far.

In the presented **publication 8**, we investigated a Miniature Shetland pony with a dwarf phenotype. This filly revealed a disproportionate short stature, a shortened neck, torso and a malformed skull. After four days, it was euthanized due to acute respiratory problems, which turned out to be a result of an aspiration pneumonia due to a soft and hard palate found in computer tomography (CT) scan. Pathological examination confirmed these findings and showed that the caudal end of the soft palate protruded into the nasal passage. Furthermore, the metacarpal, metatarsal and cannon bones were comparatively short, although no signs of chondrodysplasia could be observed patho-histologically. This individual underwent whole genome sequencing analysis in order to detect the causative variant(s) for this specific type of dwarfism and malformation. This approach allowed a fast and accurate scanning of known variants involved in dwarf types and simultaneously the investigation of novel variants, similar to the procedures in neonatal intensive care units, which use whole genome sequencing for rapid differential diagnosis of genetic disorders (Saunders et al. 2012).

We identified 7,068,170 SNPs and 923,406 indels in whole genome sequencing data of the dwarf Miniature Shetland pony. These detected variants were compared to NGS data from 25 control horses of 14 different breeds and one donkey. Filtering for variants exclusively found in the dwarf Miniature Shetland pony with predicted high or moderate effects according to SnpEff predictions (Cingolani et al. 2012), revealed 26 variants that met these criteria. In total 24 were designated as missense variants, one variant was predicted to be a disruptive in-frame deletion and one variant was a potential splice acceptor variant. Further investigation of all these variants for their potential functional effects based on sequence homology and the physical properties of amino acids using SIFT (Ng & Henikoff 2003) and PolyPhen-2 (Adzhubei et al. 2010) showed only three variants located in *activator of basal transcription 1 (ABT1), SEC31-like protein 2 (SEC31B)* and *aggrecan (ACAN)* to be deleterious and probably damaging. Validation of these three variants revealed only the *ACAN* variant to be exclusively homozygous in the affected individual and heterozygous in the mare and stallion.

ACAN was found to be a strong candidate gene, which has already been shown to be involved in the development of dwarf American Miniature horse types (Eberth 2013). Four variants were previously found to be associated with different dwarf types in American Miniature horses, which were located in the C-type lectin fold domain of ACAN, designated as D1 and D2 variant, and in the sushi domain, designated as D3 and D4 variant (Eberth 2013). The detected *ACAN* variant p.A505P in our genome data analysis for the Miniature Shetland pony was a novel so far not identified variant, also located in the C-type lectin fold domain. We assume, that this variant has a major effect on protein function and thus might affect cell adhesion and interaction of cells during endochondral ossification (Drickamer & Taylor 1993).

The results of this analysis showed that novel high impact variants causing malformations can be successfully identified by screening of whole genome sequences. To investigate further the detection capacity of NGS data for traits appearing more complex, we subsequently performed an analysis for a specific coat phenotype, the curly coat. In the presented **publication 9**, we observed curly coat in horses, particularly found in American Bashkir Curly Horses and in Missouri Foxtrotters, which was accompanied with hypotrichosis in some cases (Thomas 1990; Scott 2004). Three different curly coat phenotypes were identified: a curly coat accompanied with complete hypotrichosis, a curly coat accompanied with incomplete hypotrichosis and a curly coat without hypotrichosis. Hair fibers of horses with signs of hypotrichosis appeared to be rougher compared to those without hypotrichosis. We performed genotyping of horses showing these different phenotypes on a high density bead chip (Axiom Equine Genotyping Array 670K) to narrow down potential candidate regions. GWAS for curly versus straight coat revealed a highly significant peak on equus caballus autosome (ECA) 11 at 21,899,031-35,414,844 bp. This region harbored 26 SNPs reaching the significance threshold located in the region of 202 genes including the two keratin genes *KRT24* and *KRT25*. Further 15 keratin genes were found close by this region of association. Studies on hair structure and development confirmed that hair keratins play an important role for the assembly and maintenance of hair structure (Smith & Parry 2008). That is why keratin genes were widely considered as candidate genes for curly hair development (Kikkawa et al. 2003; Cadieu et al. 2009; Gandolfi et al. 2010; Shimomura et al. 2010; Daetwyler et al. 2014). For this reason, we investigated the region of association and the proximal keratin cluster region for candidate variants in whole genome sequencing data. Scanning of NGS data from a curly coated American Bashkir Curly Horse with complete hypotrichosis, a curly coated American Bashkir Curly Horse with incomplete hypotrichosis and a curly coated Missouri Foxtrotter without hypotrichosis compared to 27 straight-coated controls revealed six variants with high or moderate effects in these regions. After genotyping of all six variants in 148 curly and 68 straight coated horses, only three variants located in *KRT25*, *keratin associated protein 16 (KRTAP16)* and *transcription factor SP6 (SP6)* remained to be associated with investigated phenotypes. However, we could exclude *KRTAP16* variant as candidate, because we found the mutant allele of this variant in a larger validation sample set of straight-coated horses of different breeds. Thus, we concluded that *KRT25* and *SP6* missense variants were causative for curly hair in horses. Furthermore, we found *KRT25* to have an epistatic effect on *SP6* variant. Horses with one or two mutant alleles in *KRT25* variant developed curly hair and an incomplete (heterozygous genotype) or complete (homozygous mutant genotype) hypotrichosis and a rough appearance of remaining hair fibers. Those individuals with both *KRT25* and *SP6* variants also revealed curly hair with a rough surface accompanied with hypotrichosis. In contrast, horses with *SP6* variant showed curly hair but no hypotrichosis. We assumed that *KRT25* variant was masking the *SP6* effect, resulting in the development of hypotrichosis and a rough hair surface. Morphologic analysis of hair fibers by scanning electron microscopy (SEM) confirmed the rough appearance of *KRT25* mutant horses and showed the hair surfaces to be irregularly desquamated and scaly in particular in homozygous mutant individuals. The same findings were detected in horses with both *KRT25* and *SP6* mutant alleles. Individuals with an exclusive mutant allele in *SP6* revealed a regular desquamation and only slightly raised scales on hair fiber surfaces, whereas straight-coated horses were found to harbor a regular desquamation as well, but a completely smooth surface. Despite these differences in-between the phenotypes of *KRT25* and *SP6* mutant horses, we found evidence for common characteristics of curly hair in curly horses of all genotypic combinations. As previous studies on human, mice and rats postulated (Tanaka et al. 2007; Thibaut et al. 2007; Kuramoto et al. 2010), the curly hair shafts in curly horses had a polymorphic shape in contrast to the circular shape of straight hair shafts, and showed an extremely restricted medulla region. Curly hair fibers were rotated in the shafts and harbored characteristic depressions that were proposed to be responsible for the curling of hair (Thibaut et al. 2007). We supposed that these findings were a result of mutant *SP6* or mutant *KRT25* leading to a modified proliferation in hair follicles and thus an asymmetry in the hair bulb. We found evidence for such proliferative variances in studies on SP6-knockout mice that developed short curly whiskers, and in human curly hair types (Thibaut et al. 2005; Thibaut et al. 2007; Hertveldt et al. 2008). Furthermore, mutations in *KRT25* were shown to be responsible for a disarrangement of the macrofibril structure and thus a twisted appearance in human curly hair

(Thibaut et al. 2007; Ansar et al. 2015). The hair was found to be coarse and lusterless and thus was designated as woolly hair (Ansar et al. 2015). *KRT25* mutant patients additionally developed a decreased hair density (Zernov et al. 2016). These hypotrichotic signs were observed in various human patients with woolly hair, affected by different variants in other keratin genes, too (Wasif et al. 2011; Fujimoto et al. 2012).

This confirmed our findings in *KRT25* mutant curly horses with hypotrichosis. For further estimation of functional effects, we additionally performed RNA-seq in curly and straight-coated horses. No differential expression could be observed in *KRT25* or *SP6*. Nevertheless, we found significant differential expression in 20 genes involved in hair development processes located close by the region of GWAS. The two genes *KRT17* and *SOX9* were confirmed to show a significantly increased expression in curly hair in the validation analysis of 38 additionally investigated samples. Both genes were predicted to be co-expressed with *KRT25* and *SP6* according to gene network analysis (Warde-Farley et al. 2010). Furthermore, a potential interactor with *KRT25*, *the keratin gene 1 (KRT1)*, showed a significant differential expression for curly versus straight-coated horses. In the comparative analysis of hypotrichotic horses with non-hypotrichotic horses, *KRT1* and also *KRT79* revealed significant different expression levels. We found evidence that *KRT1, KRT79* and *SOX9* were involved in the regeneration of hair (Ito et al. 2005; Runkel et al. 2006; Mardaryev et al. 2011; Veniaminova et al. 2013). We concluded that these co-expressed genes might play a role in regenerative processes, which should be particularly required in hair with an irregular desquamation and scaly surface found in curly horses.

Thus, we detected two variants involved in curly hair and hypotrichosis in horses using whole genome sequencing and additionally gained knowledge about functional effects by RNA-seq analysis. Based on these data, our investigations further aimed at the curly coat phenotype found in other domestic animals. Curly hair, or woolly hair as designated in human, was identified in various species including pigs (Rhoad 1934), dogs (Cadieu et al. 2009), cats (Robinson 1969; Gandolfi et al. 2010; Gandolfi et al. 2013), cattle (Craft & Blizzard 1934; Johansson 1942; Eldridge et al. 1949), sheep (Kang et al. 2013), mice (Runkel et al. 2006; Tanaka et al. 2007), rats (Kuramoto et al. 2010), chicken (Ng et al. 2012) among others. In presented **publication 10** we investigated curly coat in Angus cattle, a breed, in which no evidence was found for a curly coat type until now (Daetwyler et al. 2014). We observed a curly-coated German Angus cow with well-defined curls in particular pronounced at neck and shoulder. Whole genome sequencing was performed for this curly-coated individual and mapped to the reference genome UMD 3.1. Comparative analysis with NGS data from 51 controls of seven cattle breeds, which were expected to express a straight coat, revealed 230 variants with predicted high or moderate effects according to SnpEff predictions (Cingolani et al. 2012), which could be exclusively found in the curly-coated cow. After scanning of all genes affected by the these variants for candidate genes known to be involved in curly or woolly coat phenotypes in mammals according to NCBI (www.ncbi.nlm.nih.gov) and OMIA (www.omia.org) database, three genes including *keratin 27* (*KRT27)*, *mitochondrially encoded NADH dehydrogenase 3 (ND3)* and *mitochondrially encoded NADH dehydrogenase 3 (ND6)* remained. Only *KRT27* variant (rs384881761) co-segregated with curly-coat phenotypes in further investigated German Angus. Genotyping of *KRT27* variant in 284 cattle of 11 different breeds confirmed the curly-coat associated variant in *KRT27* in German Angus as well as German Fleckvieh cattle. It supported the previous observation of curly coat in the German Fleckvieh population expressed by this *KRT27* (rs384881761) variant (Daetwyler et al. 2014). The frequency of this variant was low in these two breeds, whereas none of the other investigated breeds harbored a mutant allele. A neighboring variant located in *KRT24,* which was found to show a high linkage disequilibrium with *KRT27* variant, could be excluded as candidate variant for curly coat in validation analysis, because six straight-coated German Fleckvieh bulls showed a heterozygous genotype for this variant. These bulls did not harbor the *KRT27* mutant allele and thus were confirmed as straight-coated.

The results of this study revealed that curly-coat caused by a *KRT27* variant can be found in German Angus as well as German Fleckvieh cattle. Due to the origin of German Angus from crossbreeding of Aberdeen Angus with German dual-purpose breeds like German Fleckvieh (Sambraus 2011) this variant was presumably transferred into the German Angus breed. Further variants as potential candidates for curly coat in German Angus could be excluded by high throughput sequencing analysis.

All three studies in **publication 8, 9** and **10** show that whole genome data can be successfully used for the detection of variants involved in specified phenotypes. It allows a targeted screening for candidate genes and a wide-ranging detection of gene interactions. Further RNA-seq analyses provide additional information on functional effects of variants and consequences of variants on gene expression.

7 General Discussion

The present work is focused on genomics in domestic animals, providing insight into genome structure, homozygosity regions, mapping accuracy of variants, functional variant detection and evaluation of effects using NGS technology. The common key for all studies performed for this habilitation thesis is the tracking of signals for genomic changes as a result of strong selection for domestic animal populations or breeds in the course of domestication.

Diversity and signatures of selection

Domestic animals represent particularly interesting models for genome research due to their characteristic distribution into breeds or populations. Since domestication, artificial selection for specific types suitable for working, efficient use of meat or milk or even for show or competitions have highly shaped these animals (Calo et al. 1973; Outram et al. 2009; Schröder et al. 2011). As it has been demonstrated in dogs and horses, the number of genetic variants explaining the majority of phenotypic variation across breeds is comparatively low, producing markedly different breed types (Boyko et al. 2010; Petersen et al. 2013). In contrast, within breeds the phenotypic and genetic uniformity of individuals is quite high (Arranz et al. 2001; Petersen et al. 2013). Inbreeding and artificial insemination, multiple ovulation and embryo-transfer have shaped animals into characteristic and efficient breeds (Medugorac et al. 2009). Desired traits like body size, fertility or coat color were specifically selected, whereas traits under less attention received a relaxation of selection (Alberto et al. 2018).

In our studies, we identified genomic regions under strong selection for important phenotypes including fertility, variability of the immune system, metabolism, neurologic control, body size and coat structure in horses and in cattle. It confirmed that natural as well as artificial selection alter the evolutionary development of populations (Akey et al. 2010). In non-breed horses, which have not been subject to the same intensity of artificial selection and closed breeding (Warmuth et al. 2013), we found private variants affecting coding regions enriched with genes involved in metabolic processes and morphologic development (**publication 1**). This leads to the suggestion that non-breeds differ considerably from breeds due to their stronger need for survival. In feral horses is was shown that the foaling rates and survival was highly dependent on ecological range conditions, in particular on food resources (Siniff et al. 1986). In contrast, breed horses develop characteristics highly dependent on the goal of selection. We found an enrichment of genes in private variants in coding regions of Hanoverian and Arabian horses affecting neurological control and muscle contraction among others. This clearly reflects the

specialization of these breeds for performance traits as riding and show horses (Brooks et al. 2010a; Schröder et al. 2011).
Footprints of this artificial selection and inbreeding events were shown to emerge in certain regions of the genome by means of ROHs (Kim et al. 2013). These long stretches of homozygous segments were investigated in various domestic animal genomes, including cattle, pigs, sheep and dogs, in order to identify inbreeding levels, selection pressures, population structure and demographic history (Boyko et al. 2010; Bosse et al. 2012; Purfield et al. 2012; Beynon et al. 2015). In particular, intensive selection programs were shown to have increased the similarity among genomes of breeds and supported the development of ROH regions (Kim et al. 2013). In our analysis of whole genome sequencing data of horse breeds and non-breeds, a notably high total length of the genome was found to harbor ROH regions in the Thoroughbred and the two Sorraia horses (**publication 2**). This confirmed the postulated comparatively high inbreeding levels in these populations (Cunningham et al. 2001; Aberle et al. 2004). Inbreeding coefficients derived from ROH (F_{ROH}), depicting the proportion of the genome with ROH regions, were proposed to highly correlate with inbreeding coefficients derived from pedigree data (Fped) (Ferencakovic et al. 2011). However, F_{ROH} was suggested to be the most powerful method of detecting inbreeding effects, because it affects identical haplotypes over various generations that cannot be acquired by traditional pedigree-based analyses (Keller et al. 2011). F_{ROH} estimations in Holstein cattle suggested both the occurrence of bottlenecks during breed formation as well as a constant directional selection to be the cause of high inbreeding in this population (Kim et al. 2013). These findings clearly supported the assumption that selection pressures are the main drivers for inbreeding, leading to much higher F_{ROH} in domestic animals in comparison to most human populations (Curik et al. 2014; Di Gaetano et al. 2014).
In particular, artificial insemination has supported selection programs and accelerated genomic changes (Kim et al. 2013). For this reason, high fertility gained more and more importance for breeding progresses and therefore represented an essential target of indirect selection of stallions or bulls (Mackinnon et al. 1990; Colenbrander et al. 2003). Our results from investigations of selection signatures in horse breeds and high impact variants for stallion fertility (**publication 2 and 3**) confirmed that positive as well as negative fertility-related variants have a crucial influence on selection and can thus be traced back by means of footprints of selection pressures. It was postulated that important performance parameters, for example milk production, could be negatively correlated with fertility, mainly promoted by high inbreeding levels (VanRaden et al. 2004). However, in Jersey cattle is was shown that most of all detected ROH regions associated with fertility were related to higher fertility, whereas only

a small amount of ROHs was related to lower fertility (Kim et al. 2015). The genetic improvement of female fertility was suggested to be achieved by indirect selection for longevity or body condition score or direct selection of daughter pregnancy rates (Weigel 2006). These findings show that the result of selection can vary due to several interrelated factors.

The consequences of selection signatures identified by ROHs become particularly prominent, if they are related to desired key traits, whose phenotypes can be clearly defined. Such breed-defining traits like body size or coat characteristics were supposed to highlight the important role of human-directed selection on the genetic architecture of species (Boyko et al. 2010). We found evidence for this assumption in our investigations for miniature size in horses and for curly coat in horses and in cattle (**publication 4, 9 and 10**). We identified ROH regions harboring four height-associated variants in Shetland ponies of different sizes (0.28-3.34 Mb) indicating past selection pressures on miniature size reaching back to 1200-1800 years ago. In this case, strong-targeted selection in both directions for miniature as well as standard size resulted in selection signatures on size-related genes. Short ROHs indicated past eras of population processes shaping the miniature horse type, whereas longer ROHs were presumably a product of recent inbreeding similar to observations made in dog breeds (Boyko et al. 2010). Thus, the size of ROHs and its distribution among chromosomes give evidence for demographic histories of populations, recently applied in studies among various breeds including cattle and pigs (Kirin et al. 2010; Purfield et al. 2012; Herrero-Medrano et al. 2013).

However, long stretches of homozygous regions do not only provide an insight into inter-breed specific traits but also give evidence for specific phenotypes within breeds (Chan et al. 2010). A shared homozygous haplotype was observed in Portuguese water dogs, harboring two different coat types, curly coat and wavy coat, within this breed (Cadieu et al. 2009). In this region, a nonsynonymous variant in *KRT71* was identified to be causative for curly coat representing a signature of selection for both desired phenotypes, curly and wavy (Cadieu et al. 2009). In contrast to these findings, curly coat in cattle was suggested to represent a rare coat variation in Fleckvieh cattle (Daetwyler et al. 2014), in some studies even estimated as undesirable trait due to potential higher tick and parasite burden (Martinez et al. 2006; Gasparin et al. 2007). We observed this very specific coat type in German Angus individuals and assumed that it might be transmitted from Fleckvieh cattle into this breed by random selection of curly-coated individuals. As if contrary to this, however, curly coat in horses, in particular in American Bashkir Curly Horses, is highly desired and underlies strong artificial selection for this trait (Thomas 1990; Scott 2004). This is especially reasoned by the hypoallergenic potential discovered in some curly-coated horses (Mitlehner 2013). These findings propose that very

different levels of importance are attached to specific characteristics like coat phenotypes, which mainly influence the intensity of selection pressures resting on such traits.
However, consequences of selection are not only of a positive nature but can also promote unfavorable effects. Targeted selection for muscular growth in pigs was suggested to result in metabolic dysbalances within the muscle and, therefore, degenerative processes (Weiler et al. 1995). In cattle, the promotion of increased body size of Holsteins was found to show evidence for negative relation with survival and efficiency (Weiler et al. 1995). In our study, we identified a novel variant in *ACAN* causative for dwarfism accompanied with malformations in miniature-sized Miniature Shetland ponies (**publication 8**). We assumed that this type of dwarfism might be associated with extremely small body size, as it has already been suggested in dwarf American Miniature horses (Eberth 2013) and miniature cattle breeds including Dexter, Scottish Highland and Miniature Belted Galloway (Cavanagh et al. 2007; Catalina Cabrera et al. 2016; Dittmer et al. 2017). Such negative interrelations triggered by strong selection for a highly favored phenotype are usually associated with an increase of inbreeding levels (Mastrangelo et al. 2018). In particular in dog breeds, the frequent use of popular sires in closed populations and historical bottlenecks have led to a loss of diversity (Leroy 2011). In the Lundehund, a breed of a very small population size, we confirmed this suggestion and found a high frequency of individuals affected with the Lundehund syndrome (**publication 5**). This disease was probably promoted in this breed due to strong directional selection for Lundehund-specific traits like the characteristic additional toe and joint flexibility (Melis et al. 2013). The Shar-Pei breed represents a further example of indirect promotion of negative effects by targeted selection of specific phenotypes. In our study, we observed Shar-Pei dogs affected with SPAID (**publication 6 and 7**), which prominently showed a highly wrinkled skin. Evidence suggested that the gene locus for this desired characteristic skin development in Shar-Pei might be close to the causative variant for SPAID, increasing the frequent appearance of both traits. Consequently, all these findings discussed in this work illustrate the closely interlinked effects of natural and artificial selection, which were carefully traced back by means of genomic footprints.

Genome analyses- future perspectives

During the first International Symposium on Animal Genomics for Animal Health held in Paris, in October 2007, more than 250 participants from 26 countries emphasized the need to exploit the potential of new genome-enabled technologies in animal health research in order to improve health and welfare of animals worldwide (Archibald et al. 2008). The maintenance of health was proposed to be one of the biggest challenges for efficient livestock production in the next few decades (Plastow 2016). In particular, resistance to diseases, resilience, high impact variants and the incidence of malformations were outlined as essential parameters, which should be widely investigated in animal genomes (Albers et al. 1987; Thomsen et al. 2006; Fritz et al. 2013). Furthermore, it was postulated that genome analyses for population genetic aspects including diversity, inbreeding and development of breeds or non-breeds represent an essential basis for animal health, species survival and an increased animal production (Axelsson et al. 2013; Petersen et al. 2013; Daetwyler et al. 2014; Wiener et al. 2017).

In our investigations of domestic animal genomes, we demonstrated the wide-ranging and powerful use of next generation sequencing. We identified millions of variants as basis for versatile genome analyses, signatures of selection characterizing populations, high impact variants in highly selected regions and variants causative for diseases and malformations. This highlighted the potential to identify SNPs/SNVs, indels and CNVs in whole genome sequences and further functional evaluation of genome data in RNA-seq analyses. Thus, these results underline the inestimable value of high throughput technologies for research approaches. However, the advances in library preparation and bioinformatic analyses have not come to a halt yet. Recent developments of library preparation procedures have implemented kits for increasingly smaller input concentrations of nucleic acids and allowed preparations of formalin fixed tissues as well as biopsies for DNA sequencing (Malapelle et al. 2016; Wolf et al. 2016; So et al. 2018). In addition, single cell sequencing, a sequencing procedure for individual cells, more and more moved into focus for investigations of complex ecology of heterogeneous cell states and the functional evaluation of an individual cell in its microenvironment (Eberwine et al. 2014). Advanced library preparation protocols for single-cell or multi-cell DNA samples were run on high-throughput automated barcoding and library construction systems (Chromium System, 10X Genomics, Pleasanton, California) in order to preserve long-range information on genomes (Pellegrino et al. 2017). This approach particularly aims at meeting the challenges of structural variant detection and analysis of complex genomic regions (Greer et al. 2017; Spies et al. 2017). Short reads in a long-range context are recognized by the same barcode and thus are used to improve scaffolding using Supernova software (Mohr et al. 2017). These new tools point out the efforts being made to implement improved protocols for short-read sequencers in existing sequencing infrastructures (Pellegrino et al. 2017). Simultaneously, third generation sequencing technologies gain increased importance for genome research in particular for studying so far not investigated species and reconstruction of reference genomes (Bickhart et al. 2017).

Spurred on by the successes of these methods, research questions more and more focus not only on variations in the genome but also on functional effects of these variations on gene expression (Archibald et al. 2008). Translating animal genomes to phenomes, referred to as gene expression data, is proposed to be a new challenge for domestic animal research (Andersson et al. 2015). The FAANG research initiative is expected to provide comprehensive maps of functional genomic elements in animals in the future (Andersson et al. 2015). So far, initial studies on functional annotations were performed for the identification of mRNA as well as long-noncoding RNA (Koufariotis et al. 2015), micro RNA (Do et al. 2017), allele specific gene expressions (Chamberlain et al. 2015) and 3D structures of entire mammalian genomes (Koufariotis et al. 2015). In addition, further perspectives are outlined in the era of "omics", comprising not only genomics and transcriptomics but also proteomics and metabolomics to create a more complete picture of living organisms (Bhatnagar et al. 2008). Studies conducted on genomes and metabolomes in cattle postulated that the integrative strategy significantly improves the understanding of complex regulatory networks (Widmann et al. 2013).

Even more, latest developments in human health research offer an insight into the future of genome research not only for humans but also for domestic animals. In an upcoming Keystone Symposia conference, international experts from science, healthcare and industry are announced to discuss the scientific challenges of human genome sequencing and its impact on healthcare (Keystone Symposia conference in Hanover on June 4-8 2018). The focus is on the key term "precision medicine", referring to individualized treatment of a patient based on his/her own genetic, biomarker, phenotypic, or psychosocial characteristics (Jameson & Longo 2015). By means of latest technologies allowing studies on genomics, proteomics, metabolomics as well as pharmacogenomics (Collins & Varmus 2015), a totally new path will be set in all areas of medicine.

Among all these prospects in the field of "omics", NGS was found to evolve into the molecular microscope finding its way into virtually every field of biomedical research (Buermans & den Dunnen 2014). The results of present habilitation thesis provide an insight into the high potential of this new genome-enabled technology and its impact on domestic animal research in the future.

8 Summary

Exploring genomes of domestic animals using Next Generation Sequencing

Dr. med. vet. Julia Metzger

Domestic animals represent a diverse group of populations comprising millions of genetic variants in their genomes. In particular, due to their extensive diversity among breeds and high similarity within breeds, they have been shown to be a unique model for investigations of phenotypic variations.

Whole genome sequencing of highly selected horse breeds and non-breed horses kept under free range conditions revealed an average number of 3.4 million SNVs and 0.9 million indels, which could not be found in public databases. Private variants exclusively identified in non-breed horses indicated a natural selection of these horses for metabolic processes and morphogenesis, whereas breed horses harbored private variants in performance-associated genes affecting muscle contraction, neurological processes and ion/cation transport. Comparative analysis with bead chip genotyping data confirmed a high detection accuracy of next generation sequencing data of more than 95%.

Further investigations of horse breeds and non-breeds for ROH regions indicating potential signatures of selection revealed an average number of 3492 ROHs in windows of a minimum of 50 consecutive homozygous SNPs and 292 ROHs in windows of a minimum of 500 consecutive homozygous SNPs. Only three ROHs were found to be common in whole genome sequencing data of all investigated 10 individuals, harboring the *KITLG* gene. This gave evidence for a positive natural and artificial selection of breeds as well as non-breeds for reproduction traits.

Filtering of whole genome data from 11 horses for high-impact variants in genes involved in male reproductive processes resulted in 17 variants with no homozygous mutant genotype in investigated fertile stallions. After validation in further 337 fertile Hanoverian stallions, 9 high-impact variants remained and were considered as potentially deleterious factors for stallion fertility. Furthermore, a splice-site disruption variant in *NOTCH1* was shown to be associated with the de-regressed breeding values of the paternal component of the pregnancy rate per estrus (EBV-PAT). Thus, this variant was proposed to be a significant locus for Hanoverian stallion reproduction, highly affected by targeted selection for stallion fertility.

Genomic changes due to artificial selection for conformational parameters were further investigated Shetland ponies. Potential signatures of selection for extremely small body size were identified in both miniature and standard-sized ponies. NGS data revealed four synergistically interacting variants in ROH regions, which were proposed to limit the height at the withers in Shetland ponies to 87 cm (34.25 inches). We found evidence that these variants were Shetland pony-specific variants, providing a miniature size under specific genotypic combinations.

Potential signatures of selection were also identified in Lundehund dogs, investigated for the Lundehund syndrome. A missense mutation was found in whole genome sequencing data of affected Lundehund dogs in a region 2 Mb proximal to a ROH. The affected gene *LEPREL1* was estimated to be a precursor of inflammatory effects promoting the Lundehund syndrome in this breed. Similar observations were made in Shar-Pei dogs. Analyses for potential signatures of selection resulted in a large ROH region exclusively identified in whole genome sequencing data of two SPAID-affected Shar-Pei. A missense variant located in this region in *MTBP* with a predicted damaging effect was shown to be highly associated with SPAID, whereas a significant association with a previously identified CNV near *HAS2* could not be confirmed. Affected dogs revealed a characteristic Shar-Pei fever as well as other signs of inflammation including arthritis and dermatitis.

Further scans of NGS data for potential candidate mutations, successfully detected disease traits and desired phenotypes in horses and cattle. We identified a missense mutation in *ACAN* in a Miniature Shetland pony with a dwarf phenotype suggesting this variant as a novel mutation for a so far not observed dwarfism phenotype associated with malformations in Shetland ponies. Further studies on coat characteristics in horses revealed two variants associated with curly hair in *KRT25* and *SP6*. Mutant *KRT25* allele was found to be masking *SP6* effect, resulting in curly hair accompanied with hypotrichosis. This gave evidence for an epistatic effect of *KRT25* on *SP6* for the development of curly hair and hypotrichosis in horses. In cattle, whole genome sequencing data of a curly-coated German Angus compared to 51 cattle from seven different breeds resulted in a causative variant in *KRT27,* which has also been observed in curly-coated German Fleckvieh, suggesting a transmission of this variant from German Fleckvieh to German Angus.

The results of this work give an insight into the wide spectrum of genome analyses using NGS data and elucidate various genetic effects on the development of domestic animals identified as genomic footprints of selection.

9 References

Aberle, K. S., Hamann, H., Drögemüller, C. & Distl, O. (2004): Genetic diversity in German draught horse breeds compared with a group of primitive, riding and wild horses by means of microsatellite DNA markers. *Anim. Genet.* **35**: 270-7.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., et al. (2010): A method and server for predicting damaging missense mutations. *Nat. Methods* **7**: 248-9.

Akey, J. M., Ruhe, A. L., Akey, D. T., Wong, A. K., Connelly, C. F., et al. (2010): Tracking footprints of artificial selection in the dog genome. *Proceedings of the Nati Acad Sci* **107**: 1160-5.

Alam, M. J., Fatima, N., Devi, G. R., Ravins & Singh, R. K. (2012): The enhancement of stability of p53 in MTBP induced p53-MDM2 regulatory network. *Biosystems* **110**: 74-83.

Albers, G. A., Gray, G. D., Piper, L. R., Barker, J. S., Le Jambre, L. F., et al. (1987): The genetics of resistance and resilience to Haemonchus contortus infection in young merino sheep. *Int. J. Parasitol.* **17**: 1355-63.

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., et al. (2018): Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* **9**: 813.

Alkan, C., Coe, B. P. & Eichler, E. E. (2011): Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**: 363-76.

Amos, W., Driscoll, E. & Hoffman, J. I. (2011): Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proc. Biol. Sci.* **278**: 1183-8.

Anders, S. & Huber, W. (2012): Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL).*

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., et al. (2015): Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**: 57.

Andrews, S. (2010):FastQC: A quality control tool for high throughput sequence data. In: *Reference Source*, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ansar, M., Raza, S. I., Lee, K., Irfanullah, Shahi, S., et al. (2015): A homozygous missense variant in type I keratin KRT25 causes autosomal recessive woolly hair. *J. Med. Genet.* **52**: 676-80.

Antonopoulou, I., Mavrogiannis, L. A., Wilkie, A. O. & Morriss-Kay, G. M. (2004): Alx4 and Msx2 play phenotypically similar and additive roles in skull vault differentiation. *J. Anat.* **204**: 487-99.

Archibald, A., Audonnet, J.-C., Babiuk, L., Bishop, S., Gay, C., et al. (2008) Animal genomics for animal health report: critical needs, problems to be solved, potential solutions, and a roadmap for moving forward. In: *Animal Genomics for Animal Health* (pp. 407-24. Karger Publishers.

Archibald, A. L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M. A., et al. (2010): Pig genome sequence--analysis and publication strategy. *BMC Genomics* **11**: 438.

Arranz, J.-J., Bayón, Y. & San Primitivo, F. (2001): Differentiation among Spanish sheep breeds using microsatellites. *Genetics Selection Evolution* **33**: 529.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., et al. (2013): The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360-4.

Azuma, H., Yamada, Y., Shibuya-Fujiwara, N., Yamaguchi, M., Murahashi, H., et al. (2002): Functional evaluation of ex vivo expanded cord blood lymphocytes: possible use for adoptive cellular immunotherapy. *Exp. Hematol.* **30**: 346-51.

Bactrian Camels Genome, S., Analysis, C., Jirimutu, Wang, Z., Ding, G., et al. (2012): Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* **3**: 1202.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., et al. (2012): The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587-93.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., et al. (2015): Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences* **112**: 5473-8.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., et al. (2014): PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10**: e1003440.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., et al. (2008): Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53.

Beynon, S. E., Slavov, G. T., Farre, M., Sunduimijid, B., Waddams, K., et al. (2015): Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genet* **16**: 65.

Bhatnagar, D., Rajasekaran, K., Payne, G., Brown, R., Yu, J., et al. (2008): The'omics' tools: genomics, proteomics, metabolomics and their potential for solving the aflatoxin contamination problem. *World Mycotoxin Journal* **1**: 3-12.

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., et al. (2012): Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.*

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., et al. (2017): Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**: 643-50.

Binns, M. M., Boehler, D. A., Bailey, E., Lear, T. L., Cardwell, J. M., et al. (2012): Inbreeding in the Thoroughbred horse. *Anim. Genet.* **43**: 340-2.

Blume-Jensen, P., Jiang, G., Hyman, R., Lee, K. F., O'Gorman, S., et al. (2000): Kit/stem cell factor receptor-induced activation of phosphatidylinositol 3'-kinase is essential for male fertility. *Nat. Genet.* **24**: 157-62.

Bosse, M., Megens, H. J., Madsen, O., Paudel, Y., Frantz, L. A., et al. (2012): Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet.* **8**: e1003100.

Bovine Genome, S., Analysis, C., Elsik, C. G., Tellam, R. L., Worley, K. C., et al. (2009): The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522-8.

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., et al. (2010): A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* **8**: e1000451.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., et al. (2008): The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**: 1146-53.

Brooks, S. A., Gabreski, N., Miller, D., Brisbin, A., Brown, H. E., et al. (2010a): Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for Lavender Foal Syndrome. *PLoS Genet.* **6**: 650.

Brooks, S. A., Makvandi-Nejad, S., Chu, E., Allen, J. J., Streeter, C., et al. (2010b): Morphological variation in the horse: defining complex traits of body size and shape. *Anim. Genet.* **41 (suppl. 2)**: 159-65.

Buermans, H. P. & den Dunnen, J. T. (2014): Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* **1842**: 1932-41.

Burren, A., Neuditschko, M., Signer-Hasler, H., Frischknecht, M., Reber, I., et al. (2016): Genetic diversity analyses reveal first insights into breed-specific selection signatures within Swiss goat breeds. *Anim. Genet.* **47**: 727-39.

Cadieu, E., Neff, M. W., Quignon, P., Walsh, K., Chase, K., et al. (2009): Coat variation in the domestic dog is governed by variants in three genes. *Science* **326**: 150-3.

Cai, Q., Long, J., Lu, W., Qu, S., Wen, W., et al. (2011): Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum. Mol. Genet.* **20**: 4991-9.

Calo, L., McDowell, R., VanVleck, L. D. & Miller, P. (1973): Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. *J. Anim. Sci.* **37**: 676-82.

Canavez, F. C., Luche, D. D., Stothard, P., Leite, K. R., Sousa-Canavez, J. M., et al. (2012): Genome sequence and assembly of Bos indicus. *J. Hered.* **103**: 342-8.

Capon, F., Allen, M. H., Ameen, M., Burden, A. D., Tillman, D., et al. (2004): A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.* **13**: 2361-8.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., et al. (1999): Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**: 231-8.

Carneiro, M., Rubin, C. J., Di Palma, F., Albert, F. W., Alfoldi, J., et al. (2014): Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* **345**: 1074-9.

Catalina Cabrera, L., McNabb, B. R., Woods, S. E., Cartoceti, A. N. & Busch, R. C. (2016): Hydrops associated with chondrodysplasia of the fetus in a miniature Scottish Highland cow. *J. Am. Vet. Med. Assoc.* **248**: 552-6.

Cavanagh, J. A., Tammen, I., Windsor, P. A., Bateman, J. F., Savarirayan, R., et al. (2007): Bulldog dwarfism in Dexter cattle is caused by mutations in ACAN. *Mamm. Genome* **18**: 808-14.

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., et al. (2015a): Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608.

Chaisson, M. J., Wilson, R. K. & Eichler, E. E. (2015b): Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**: 627-40.

Chamberlain, A. J., Vander Jagt, C. J., Hayes, B. J., Khansefid, M., Marett, L. C., et al. (2015): Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**: 993.

Chan, E. K., Nagaraj, S. H. & Reverter, A. (2010): The evolution of tropical adaptation: comparing taurine and zebu cattle. *Anim. Genet.* **41**: 467-77.

Chasman, D. & Adams, R. M. (2001): Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**: 683-706.

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., et al. (2009): Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* **26**: 1523-31.

Chen, L., Jin, P. & Qin, Z. S. (2016): DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol.* **17**: 252.

Chen, R., Davydov, E. V., Sirota, M. & Butte, A. J. (2010): Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* **5**: e13574.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., et al. (2017): novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods* **14**: 65-7.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., et al. (2012): A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

Colenbrander, B., Gadella, B. M. & Stout, T. A. (2003): The predictive value of semen analysis in the evaluation of stallion fertility. *Reprod. Domest. Anim.* **38**: 305-11.

Collins, F. S. & Varmus, H. (2015): A new initiative on precision medicine. *N. Engl. J. Med.* **372**: 793-5.

Colquitt, B. M., Mets, D. G. & Brainard, M. S. (2018): Draft genome assembly of the Bengalese finch, Lonchura striata domestica, a model for motor skill variability and learning. *GigaScience*.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., et al. (2016): A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**: 13.

Consortium, E. P. (2012): An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57.

Cortazzo, P., Cerveñansky, C., Marín, M., Reiss, C., Ehrlich, R., et al. (2002): Silent mutations affect in vivo protein folding in Escherichia coli. *Biochem. Biophys. Res. Commun.* **293**: 537-41.

Craft, W. & Blizzard, W. (1934): The inheritance of semi-hairlessness in cattle. *J. Hered.* **25**: 385-90.

Cunningham, E., Dooley, J., Splan, R. & Bradley, D. (2001): Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Anim. Genet.* **32**: 360-4.

Curik, I., Ferenčaković, M. & Sölkner, J. (2014): Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livestock Science* **166**: 26-34.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., et al. (2014): Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* **46**: 858-65.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., et al. (2010): Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* **8**.

Darki, F., Peyrard-Janvid, M., Matsson, H., Kere, J. & Klingberg, T. (2012): Three dyslexia susceptibility genes, DYX1C1, DCDC2, and KIAA0319, affect temporo-parietal white matter structure. *Biol. Psychiatry* **72**: 671-6.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., et al. (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**: 491-8.

Di Gaetano, C., Fiorito, G., Ortu, M. F., Rosa, F., Guarrera, S., et al. (2014): Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection. *PLoS ONE* **9**: e91237.

Dittmer, K., Thompson, K. & Hogan, T. (2017): Severe generalised chondrodysplasia in miniature cattle breeds. *N. Z. Vet. J.* **65**: 282-3.

Do, D. N., Li, R., Dudemaine, P. L. & Ibeagha-Awemu, E. M. (2017): MicroRNA roles in signalling during lactation: an insight from differential expression, time course and pathway analyses of deep sequence data. *Sci Rep* **7**: 44605.

Doan, R., Cohen, N. D., Sawyer, J., Ghaffari, N., Johnson, C. D., et al. (2012): Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. *BMC Genomics* **13**: 78.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., et al. (2013): STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. (2008): Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**: e105.

Drickamer, K. & Taylor, M. E. (1993): Biology of animal lectins. *Annu. Rev. Cell Biol.* **9**: 237-64.

Eberth, J. E. (2013):Chondrodysplasia-Like Dwarfism in the Miniature Horse. University of Kentucky, Kentucky.

Eberwine, J., Sul, J. Y., Bartfai, T. & Kim, J. (2014): The promise of single-cell sequencing. *Nat. Methods* **11**: 25-7.

Eck, S. H., Benet-Pagès, A., Flisikowski, K., Meitinger, T., Fries, R., et al. (2009): Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. *Genome Biol.* **10**: R82.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., et al. (2009): Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-8.

Eldridge, F. E., Atkeson, F. & Ibsen, H. L. (1949): Inheritance of a karakul-type curliness in the hair of Ayrshire cattle. *J. Hered.* **40**: 205-14.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., et al. (2012): Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**: e47768.

Ferencakovic, M., Hamzic, E., Gredler, B., Curik, I. & Sölkner, J. (2011): Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. *Agriculturae Conspectus Scientificus (ACS)* **76**: 325-9.

Feuk, L., Carson, A. R. & Scherer, S. W. (2006): Structural variation in the human genome. *Nat. Rev. Genet.* **7**: 85-97.

Fitak, R. R., Mohandesan, E., Corander, J. & Burger, P. A. (2016): The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Molecular ecology resources* **16**: 314-24.

FlesjÅ, K. & Yri, T. (1977): Protein-losing enteropathy in the Lundehund. *J. Small Anim. Pract.* **18**: 11-23.

Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G., McHardy, A. C., et al. (2016): Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* **6**: 25373.

Frankl-Vilches, C., Kuhl, H., Werber, M., Klages, S., Kerick, M., et al. (2015): Using the canary genome to decipher the evolution of hormone-sensitive gene regulation in seasonal singing birds. *Genome Biol.* **16**: 19.

Frischknecht, M., Jagannathan, V., Plattet, P., Neuditschko, M., Signer-Hasler, H., et al. (2015): A Non-Synonymous HMGA2 Variant Decreases Height in Shetland Ponies and Other Small Horses. *PLoS ONE* **10**: e0140749.

Fritz, S., Capitan, A., Djari, A., Rodriguez, S. C., Barbat, A., et al. (2013): Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. *PLoS ONE* **8**: e65550.

Fujimoto, A., Farooq, M., Fujikawa, H., Inoue, A., Ohyama, M., et al. (2012): A missense mutation within the helix initiation motif of the keratin K71 gene underlies autosomal dominant woolly hair/hypotrichosis. *J. Invest. Dermatol.* **132**: 2342-9.

Fujiwara, Y., Matsumoto, H., Akiyama, K., Srivastava, A., Chikushi, M., et al. (2015): An ENU-induced mutation in the mouse Rnf212 gene is associated with male meiotic failure and infertility. *Reproduction* **149**: 67-74.

Gandolfi, B., Alhaddad, H., Joslin, S. E., Khan, R., Filler, S., et al. (2013): A splice variant in KRT71 is associated with curly coat phenotype of Selkirk Rex cats. *Sci. Rep-UK* **3**.

Gandolfi, B., Outerbridge, C. A., Beresford, L. G., Myers, J. A., Pimentel, M., et al. (2010): The naked truth: Sphynx and Devon Rex cat breed mutations in KRT71. *Mamm. Genome* **21**: 509-15.

Gasparin, G., Miyata, M., Coutinho, L., Martinez, M., Teodoro, R., et al. (2007): Mapping of quantitative trait loci controlling tick [Riphicephalus (Boophilus) microplus] resistance on bovine chromosomes 5, 7 and 14. *Anim. Genet.* **38**: 453-9.

Ghoneim, D. H., Myers, J. R., Tuttle, E. & Paciorkowski, A. R. (2014): Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res. Notes* **7**: 864.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., et al. (2004): Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493-521.

Gordon, D., Huddleston, J., Chaisson, M. J., Hill, C. M., Kronenberg, Z. N., et al. (2016): Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Gottschalk, M., Metzger, J., Martinsson, G., Sieme, H. & Distl, O. (2016): Genome-wide association study for semen quality traits in German Warmblood stallions. *Anim. Reprod. Sci.* **171**: 81-6.

Greer, S. U., Nadauld, L. D., Lau, B. T., Chen, J., Wood-Bouwens, C., et al. (2017): Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med.* **9**: 57.

Hamann, H., Jude, R., Sieme, H., Mertens, U., Töpfer-Petersen, E., et al. (2007): A polymorphism within the equine CRISP3 gene is associated with stallion fertility in Hanoverian warmblood horses. *Anim. Genet.* **38**: 259-64.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., et al. (2009): Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**: R32.

Herrero-Medrano, J. M., Megens, H. J., Groenen, M. A., Ramis, G., Bosse, M., et al. (2013): Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genet* **14**: 106.

Hertveldt, V., Louryan, S., Van Reeth, T., Dreze, P., Van Vooren, P., et al. (2008): The development of several organs and appendages is impaired in mice lacking Sp6. *Dev. Dyn.* **237**: 883-92.

Hu, M., Yu, J., Taylor, J. M., Chinnaiyan, A. M. & Qin, Z. S. (2010): On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.* **38**: 2154-67.

Huang da, W., Sherman, B. T. & Lempicki, R. A. (2009a): Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**: 1-13.

Huang da, W., Sherman, B. T. & Lempicki, R. A. (2009b): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**: 44-57.

International Chicken Genome Sequencing, C. (2004): Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.

International Sheep Genomics, C., Archibald, A. L., Cockett, N. E., Dalrymple, B. P., Faraut, T., et al. (2010): The sheep genome reference sequence: a work in progress. *Anim. Genet.* **41**: 449-53.

Ishibashi, K., Kanno, E., Itoh, T. & Fukuda, M. (2009): Identification and characterization of a novel Tre2/Bub2/Cdc16 (TBC) protein that possesses Rab3A GAP activity. *Genes Cells* **14**: 41-52.

Ito, M., Liu, Y., Yang, Z., Nguyen, J., Liang, F., et al. (2005): Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. *Nat. Med.* **11**: 1351-4.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., et al. (2018): Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*

Jameson, J. L. & Longo, D. L. (2015): Precision medicine--personalized, problematic, and promising. *N. Engl. J. Med.* **372**: 2229-34.

Ji, P., Liu, G., Xu, J., Wang, X., Li, J., et al. (2012): Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. *PLoS ONE* **7**: e35152.

Jiang, Y., Wang, Y. & Brudno, M. (2012): PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28**: 2576-83.

Johansson, I. (1942): Reduced phalanges and curly coat: Two mutant characters in native swedish cattle. *Hereditas* **28**: 278-88.

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., et al. (2006): Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences* **103**: 19635-40.

Kang, X., Liu, G., Liu, Y., Xu, Q., Zhang, M., et al. (2013): Transcriptome profile at different physiological stages reveals potential mode for curly fleece in Chinese tan sheep. *PLoS ONE* **8**: e71763.

Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. (1996): Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences* **93**: 13770-3.

Keller, M. C., Visscher, P. M. & Goddard, M. E. (2011): Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189**: 237-49.

Kikkawa, Y., Oyama, A., Ishii, R., Miura, I., Amano, T., et al. (2003): A small deletion hotspot in the type II keratin gene mK6irs1/Krt2-6g on mouse chromosome 15, a candidate for causing the wavy hair of the caracul (Ca) mutation. *Genetics* **165**: 721-33.

Kim, E.-S., Sonstegard, T. S., Van Tassell, C. P., Wiggans, G. & Rothschild, M. F. (2015): The relationship between runs of homozygosity and inbreeding in Jersey cattle under selection. *PLoS ONE* **10**: e0129967.

Kim, E. S., Cole, J. B., Huson, H., Wiggans, G. R., Van Tassell, C. P., et al. (2013): Effect of artificial selection on runs of homozygosity in u.s. Holstein cattle. *PLoS ONE* **8**: e80813.

Kim, H., Schmidt, C. J., Decker, K. S. & Emara, M. G. (2003): A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. *Anim. Genet.* **34**: 249-54.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., et al. (2007): A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**: 525-8.

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., et al. (2010): Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**: e13996.

Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. & Seelow, D. (2011): Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS ONE* **6**: e28240.

Koenen, E., Aldridge, L. & Philipsson, J. (2004): An overview of breeding objectives for warmblood sport horses. *Livestock Science* **88**: 77-84.

Kolbjørnsen, Ø., Press, C. M. & Landsverk, T. (1994): Gastropathies in the Lundehund. *APMIS* **102**: 647-61.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., et al. (2007): Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420-6.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., et al. (2012): Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**: 693-700.

Koufariotis, L. T., Chen, Y. P., Chamberlain, A., Vander Jagt, C. & Hayes, B. J. (2015): A catalogue of novel bovine long noncoding RNA across 18 tissues. *PLoS ONE* **10**: e0141225.

Kumar, P., Henikoff, S. & Ng, P. C. (2009): Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**: 1073-81.

Künstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., et al. (2016): The genome of the Trinidadian guppy, Poecilia reticulata, and variation in the Guanapo population. *PLoS ONE* **11**: e0169087.

Kuntz, R., Kubalek, C., Ruf, T., Tataruch, F. & Arnold, W. (2006): Seasonal adjustment of energy budget in a large wild mammal, the Przewalski horse (Equus ferus przewalskii) I. Energy intake. *J. Exp. Biol.* **209**: 4557-65.

Kuramoto, T., Hirano, R., Kuwamura, M. & Serikawa, T. (2010): Identification of the rat Rex mutation as a 7-bp deletion at splicing acceptor site of the Krt71 gene. *J. Vet. Med. Sci.* **72**: 909-12.

Langmead, B. & Salzberg, S. L. (2012): Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**: 357-9.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., et al. (2014): Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*: 006395.

Lee, M. A., Keane, O. M., Glass, B. C., Manley, T. R., Cullen, N. G., et al. (2006): Establishment of a pipeline to analyse non-synonymous SNPs in Bos taurus. *BMC Genomics* **7**: 298.

Leroy, G. (2011): Genetic diversity, inbreeding and breeding practices in dogs: results from pedigree analyses. *Vet. J.* **189**: 177-82.

Li, B. & Dewey, C. N. (2011): RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li, H. (2015): FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**: 3694-6.

Li, H. & Durbin, R. (2009): Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al. (2009a): The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-9.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., et al. (2009b): SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**: 1124-32.

Liang, Y., Dong, Y., Zhao, J. & Li, W. (2013): YES1 activation elicited by heat stress is anti-apoptotic in mouse pachytene spermatocytes. *Biol. Reprod.*: biolreprod. 113.112235.

Librado, P., Gamba, C., Gaunitz, C., Der Sarkissian, C., Pruvost, M., et al. (2017): Ancient genomic changes associated with domestication of the horse. *Science* **356**: 442-5.

Light, N., Adoue, V., Ge, B., Chen, S. H., Kwan, T., et al. (2014): Interrogation of allelic chromatin states in human cells by high-density ChIP-genotyping. *Epigenetics* **9**: 1238-51.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., et al. (2005): Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803-19.

Luis, C., Cothran, E. G. & Oom Mdo, M. (2007): Inbreeding and genetic structure in the endangered Sorraia horse breed: implications for its conservation and management. *J. Hered.* **98**: 232-7.

MacEachern, S., McEwan, J., Mather, A., McCulloch, A., Sunnucks, P., et al. (2006): Testing the neutral theory of molecular evolution using genomic data: a comparison of the human and bovine transcriptome. *Genetics Selection Evolution* **38**: 321.

Mackinnon, M. J., Taylor, J. F. & Hetzel, D. J. (1990): Genetic variation and covariation in beef cow and bull fertility. *J. Anim. Sci.* **68**: 1208-14.

MacLeod, I. M., Larkin, D. M., Lewin, H. A., Hayes, B. J. & Goddard, M. E. (2013): Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol. Biol. Evol.* **30**: 2209-23.

Makvandi-Nejad, S., Hoffman, G. E., Allen, J. J., Chu, E., Gu, E., et al. (2012): Four Loci explain 83% of size variation in the horse. *PLoS ONE* **7**: e39929.

Malapelle, U., Pisapia, P., Rocco, D., Smeraglio, R., di Spirito, M., et al. (2016): Next generation sequencing techniques in liquid biopsy: focus on non-small cell lung cancer patients. *Translational lung cancer research* **5**: 505.

Manukyan, G., Aminov, R., Hakobyan, G. & Davtyan, T. (2015): Accelerated apoptosis of neutrophils in familial mediterranean Fever. *Frontiers in immunology* **6**: 239.

Mardaryev, A. N., Meier, N., Poterlowicz, K., Sharov, A. A., Sharova, T. Y., et al. (2011): Lhx2 differentially regulates Sox9, Tcf4 and Lgr5 in hair follicle stem cells to promote epidermal regeneration after injury. *Development* **138**: 4843-52.

Martinez, M. L., Machado, M. A., Nascimento, C. S., Silva, M. V., Teodoro, R. L., et al. (2006): Association of BoLA-DRB3.2 alleles with tick (Boophilus microplus) resistance in cattle. *Genet. Mol. Res.* **5**: 513-24.

Mastrangelo, S., Sardina, M. T., Tolone, M., Di Gerlando, R., Sutera, A. M., et al. (2018): Genome-wide identification of runs of homozygosity islands and associated genes in local dairy cattle breeds. *Animal*: 1-9.

McGovern, D. P., Hysi, P., Ahmad, T., van Heel, D. A., Moffatt, M. F., et al. (2005): Association between a complex insertion/deletion polymorphism in NOD1 (CARD4) and susceptibility to inflammatory bowel disease. *Hum. Mol. Genet.* **14**: 1245-50.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., et al. (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**: 1297-303.

Medugorac, I., Medugorac, A., Russ, I., Veit-Kensch, C. E., Taberlet, P., et al. (2009): Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Mol. Ecol.* **18**: 3394-410.

Medvedev, P., Stanciu, M. & Brudno, M. (2009): Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6**: S13-20.

Melis, C., Borg, Å. A., Espelien, I. S. & Jensen, H. (2013): Low neutral genetic variability in a specialist puffin hunter: the Norwegian Lundehund. *Anim. Genet.* **44**: 348-51.

Metzker, M. L. (2010): Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**: 31-46.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., et al. (2004): The genome sequence of silkworm, Bombyx mori. *DNA Res.* **11**: 27-35.

Mitlehner, W. (2013): Allergy against horses: Are curly horses an alternative for horse-allergic riders? *Allergo J* **22**: 244-51.

Mohr, D. W., Naguib, A., Weisenfeld, N., Kumar, V., Shah, P., et al. (2017): Improved de novo Genome Assembly: Synthetic long read sequencing combined with optical mapping produce a high quality mammalian genome at relatively low cost. *bioRxiv*: 128348.

Molra C. Reeve, S. B. (2011) The breeds. In: *The Original Horse Bible: The Definitive Source for All Things Horse* (ed. by Stein JS). Bowtie Press, Irivine, California.

Mouse Genome Sequencing, C., Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., et al. (2002): Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-62.

Mulay, S. R., Thomasova, D., Ryu, M. & Anders, H.-J. (2012): MDM2 (murine double minute-2) links inflammation and tubular cell healing during acute kidney injury in mice. *Kidney Int.* **81**: 1199-211.

Mullinger, A. M. & Johnson, R. T. (1980): Packing DNA into chromosomes. *J. Cell Sci.* **46**: 61-86.

Narlikar, L. & Ovcharenko, I. (2009): Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomic. Proteomic.* **8**: 215-30.

Narzisi, G., O'rawe, J. A., Iossifov, I., Fang, H., Lee, Y.-h., et al. (2014): Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**: 1033.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., et al. (2012): An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**: 83-90.

Ng, C. S., Wu, P., Foley, J., Foley, A., McDonald, M. L., et al. (2012): The chicken frizzle feather is due to an alpha-keratin (KRT75) mutation that causes a defective rachis. *PLoS Genet.* **8**: e1002748.

Ng, P. C. & Henikoff, S. (2003): SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**: 3812-4.

Nijkamp, J. F., van den Broek, M. A., Geertman, J. M., Reinders, M. J., Daran, J. M., et al. (2012): De novo detection of copy number variation by co-assembly. *Bioinformatics* **28**: 3195-202.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J. & Yuhki, N. (2002): The Feline Genome Project. *Annu. Rev. Genet.* **36**: 657-86.

Olsson, M., Meadows, J. R., Truve, K., Rosengren Pielberg, G., Puppo, F., et al. (2011): A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genet.* **7**: e1001332.

Olsson, M., Tintle, L., Kierczak, M., Perloski, M., Tonomura, N., et al. (2013): Thorough investigation of a canine autoinflammatory disease (AID) confirms one main risk locus and suggests a modifier locus for amyloidosis. *PLoS ONE* **8**: e75242.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., et al. (2013): Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**: 74-8.

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., et al. (2009): The earliest horse harnessing and milking. *Science* **323**: 1332-5.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., et al. (2014): A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15**: 256-78.

Pausch, H., Kolle, S., Wurmser, C., Schwarzenbacher, H., Emmerling, R., et al. (2014): A nonsense mutation in TMEM95 encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. *PLoS Genet.* **10**: e1004044.

Pellegrino, R., Benway, M., Kocjan, P., Price, A., Kao, C., et al. (2017):High-throughput automation of the 10x Genomics® Chromium™ workflow for linked-read whole exome sequencing and a targeted lynch syndrome panel. AACR.

Peng, X., Alfoldi, J., Gori, K., Eisfeld, A. J., Tyler, S. R., et al. (2014): The draft genome sequence of the ferret (Mustela putorius furo) facilitates study of human respiratory disease. *Nat. Biotechnol.* **32**: 1250-5.

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., et al. (2013): Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.* **9**: e1003211.

Pfahler, S. & Distl, O. (2015): Effective Population Size, Extended Linkage Disequilibrium and Signatures of Selection in the Rare Dog Breed Lundehund. *PLoS ONE* **10**: e0122680.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., et al. (1998): High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207-11.

Pirooznia, M., Goes, F. S. & Zandi, P. P. (2015): Whole-genome CNV analysis: advances in computational approaches. *Frontiers in genetics* **6**: 138.

Plastow, G. S. (2016): Genomics to benefit livestock production: improving animal health. *Revista Brasileira de Zootecnia* **45**: 349-54.

Pontius, J. U., Mullikin, J. C., Smith, D. R., Agencourt Sequencing, T., Lindblad-Toh, K., et al. (2007): Initial sequence and comparative analysis of the cat genome. *Genome Res.* **17**: 1675-89.

Purfield, D. C., Berry, D. P., McParland, S. & Bradley, D. G. (2012): Runs of homozygosity and population history in cattle. *BMC Genet* **13**: 70.

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., et al. (2011): Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* **12**: 318.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., et al. (2012): DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333-i9.

Reinartz, S., Mohwinkel, H., Surie, C., Hellige, M., Feige, K., et al. (2017): Germline mutation within COL2A1 associated with lethal chondrodysplasia in a polled Holstein family. *BMC Genomics* **18**: 762.

Reuter, J. A., Spacek, D. V. & Snyder, M. P. (2015): High-throughput sequencing technologies. *Mol. Cell* **58**: 586-97.

Rhoad, A. (1934): Woolly hair in swine. *J. Hered.* **25**: 371-5.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., et al. (2014): Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**: 912.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-40.

Robinson, R. (1969): Devon rex—A third rexoid coat mutant in the cat. *Genetica* **40**: 597-9.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., et al. (2011): An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348.

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., et al. (2010): Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587-91.

Runkel, F., Klaften, M., Koch, K., Böhnert, V., Büssow, H., et al. (2006): Morphologic and molecular characterization of two novel Krt71 (Krt2-6g) mutations: Krt71 rco12 and Krt71 rco13. *Mamm. Genome* **17**: 1172-82.

Runyan, C., Schaible, K., Molyneaux, K., Wang, Z., Levin, L., et al. (2006): Steel factor controls midline cell death of primordial germ cells and is essential for their normal proliferation and migration. *Development* **133**: 4861-9.

Sambraus, H. H. (2011) *Farbatlas Nutztierrassen*. Ulmer.

Saunders, C. J., Miller, N. A., Soden, S. E., Dinwiddie, D. L., Noll, A., et al. (2012): Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**: 154ra35.

Schadt, E. E., Turner, S. & Kasarskis, A. (2010): A window into third-generation sequencing. *Hum. Mol. Genet.* **19**: R227-R40.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., et al. (2010): Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036-40.

Schmieder, R. & Edwards, R. (2011): Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863-4.

Schneider, G. F. & Dekker, C. (2012): DNA sequencing with nanopores. *Nat. Biotechnol.* **30**: 326.

Schook, L. B., Beever, J. E., Rogers, J., Humphray, S., Archibald, A., et al. (2005): Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comp Funct Genomics* **6**: 251-5.

Schröder, W., Klostermann, A., Stock, K. F. & Distl, O. (2011): A genome-wide association study for quantitative trait loci of show-jumping in Hanoverian warmblood horses. *Anim. Genet.* **doi:10.1111/j.1365-2052.2011.02265.x**.

Scott, D. (2004): Skin of the neck, mane and tail of the curly horse. *Equine Veterinary Education* **16**: 201-6.

Seemann, S. E., Anthon, C., Palasca, O. & Gorodkin, J. (2015): Quality assessment of domesticated animal genome assemblies. *Bioinformatics and Biology insights* **9**: BBI. S29333.

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., Pan, H., et al. (2013): Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339**: 1063-7.

Shearin, A. L. & Ostrander, E. A. (2010): Leading the way: canine models of genomics and disease. *Dis. Model. Mech.* **3**: 27-34.

Shimomura, Y., Wajid, M., Petukhova, L., Kurban, M. & Christiano, A. M. (2010): Autosomal-dominant woolly hair resulting from disruption of keratin 74 (KRT74), a potential determinant of human hair texture. *Am. J. Hum. Genet* **86**: 632-8.

Siniff, D. B., Tester, J. R. & McMahon, G. L. (1986): Foaling rate and survival of feral horses in western Nevada. *Journal of Range Management* **39**: 296-7.

Smith, T. A. & Parry, D. A. (2008): Three-dimensional modelling of interchain sequence similarities and differences in the coiled-coil segments of keratin intermediate filament heterodimers highlight features important in assembly. *J. Struct. Biol.* **162**: 139-51.

So, A. P., Vilborg, A., Bouhlal, Y., Koehler, R. T., Grimes, S. M., et al. (2018): A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. *NPJ genomic medicine* **3**: 2.

Sonstegard, T. S., Cole, J. B., VanRaden, P. M., Van Tassell, C. P., Null, D. J., et al. (2013): Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PLoS ONE* **8**: e54872.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., et al. (2017): Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* **14**: 915.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., et al. (2002): Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **99**: 4465-70.

Szpiech, Z. A., Xu, J., Pemberton, T. J., Peng, W., Zollner, S., et al. (2013): Long runs of homozygosity are enriched for deleterious variation. *Am. J. Hum. Genet.* **93**: 90-102.

Tanaka, S., Miura, I., Yoshiki, A., Kato, Y., Yokoyama, H., et al. (2007): Mutations in the helix termination motif of mouse type I IRS keratin genes impair the assembly of keratin intermediate filament. *Genomics* **90**: 703-11.

Thibaut, S., Barbarat, P., Leroy, F. & Bernard, B. A. (2007): Human hair keratin network and curvature. *Int. J. Dermatol.* **46 Suppl 1**: 7-10.

Thibaut, S., Gaillard, O., Bouhanna, P., Cannell, D. & Bernard, B. (2005): Human hair shape is programmed from the bulb. *Br. J. Dermatol.* **152**: 632-8.

Thomas, S. (1990) The Curly Horse identification project of the CS fund conservancy (a case study). In: *Genetic conservation of domestic livestock.* (ed. by Alderson L), pp. 154-9. CAB International, Wallingford, UK.

Thomsen, B., Horn, P., Panitz, F., Bendixen, E., Petersen, A. H., et al. (2006): A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. *Genome Res.* **16**: 97-105.

Trapnell, C., Pachter, L. & Salzberg, S. L. (2009): TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-11.

Tuggle, C. K., Giuffra, E., White, S. N., Clarke, L., Zhou, H., et al. (2016): GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. *Anim. Genet.* **47**: 528-33.

Van Bergen, H. & Van Arendonk, J. (1993): Genetic parameters for linear type traits in Shetland Ponies. *Livest. Prod. Sci.* **36**: 273-84.

van Binsbergen, R., Bink, M., Calus, M., van Eeuwijk, F. A., Hayes, B. J., et al. (2014): Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **46**: 41.

VanRaden, P. M., Sanders, A. H., Tooker, M. E., Miller, R. H., Norman, H. D., et al. (2004): Development of a national genetic evaluation for cow fertility. *J. Dairy Sci.* **87**: 2285-92.

Veniaminova, N. A., Vagnozzi, A. N., Kopinke, D., Do, T. T., Murtaugh, L. C., et al. (2013): Keratin 79 identifies a novel population of migratory epithelial cells that initiates hair canal morphogenesis and regeneration. *Development* **140**: 4870-80.

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., et al. (2009): Genome sequence, comparative analysis, and population genetics of the domestic horse. *Sci* **326**: 865-7.

Wang, C., Wang, H., Zhang, Y., Tang, Z., Li, K., et al. (2015): Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. *Molecular ecology resources* **15**: 414-24.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., et al. (2010): The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**: W214-W20.

Warmuth, V., Manica, A., Eriksson, A., Barker, G. & Bower, M. (2013): Autosomal genetic diversity in non-breed horses from eastern Eurasia provides insights into historical population movements. *Anim. Genet.* **44**: 53-61.

Wasif, N., Naqvi, S. K., Basit, S., Ali, N., Ansar, M., et al. (2011): Novel mutations in the keratin-74 (KRT74) gene underlie autosomal dominant woolly hair/hypotrichosis in Pakistani families. *Hum. Genet.* **129**: 419-24.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., et al. (2002): Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**: 854-62.

Weigel, K. A. (2006): Prospects for improving reproductive performance through genetic selection. *Anim. Reprod. Sci.* **96**: 323-30.

Weiler, U., Appell, H. J., Kremser, M., Hofacker, S. & Claus, R. (1995): Consequences of selection on muscle composition. A comparative study on gracilis muscle in wild and domestic pigs. *Anat. Histol. Embryol.* **24**: 77-80.

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., et al. (2006): Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**: 931-49.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., et al. (2008): The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872.

Widmann, P., Reverter, A., Fortes, M. R., Weikard, R., Suhre, K., et al. (2013): A systems biology approach using metabolomic data reveals genes and pathways interacting to modulate divergent growth in cattle. *BMC Genomics* **14**: 798.

Wiener, P., Sanchez-Molano, E., Clements, D. N., Woolliams, J. A., Haskell, M. J., et al. (2017): Genomic data illuminates demography, genetic structure and selection of a popular dog breed. *BMC Genomics* **18**: 609.

Winchester, L., Yau, C. & Ragoussis, J. (2009): Comparing CNV detection methods for SNP arrays. *Brief. Funct. Genomic. Proteomic.* **8**: 353-66.

Wolf, K., O'Neil, D., Schroeer, S. & Fang, N. (2016): Next-Generation Sequencing Library Preparation from FFPE Tissue Samples. *Curr. Protoc. Mol. Biol.*: 7.24. 1-7.. 14.

Wu, J. Y. (1982): Purification and characterization of cysteic acid and cysteine sulfinic acid decarboxylase and L-glutamate decarboxylase from bovine brain. *Proc. Natl. Acad. Sci. U. S. A.* **79**: 4270-4.

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., et al. (2004): A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science* **306**: 1937-40.

Yang, R., Van Etten, J. L. & Dehm, S. M. (2018): Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics* **19**: 270.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. (2009): Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-71.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. (2009): Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**: 1586-92.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., et al. (2014): A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355-64.

Zernov, N. V., Skoblov, M. Y., Marakhonov, A. V., Shimomura, Y., Vasilyeva, T. A., et al. (2016): Autosomal Recessive Hypotrichosis with Woolly Hair Caused by a Mutation in the Keratin 25 Gene Expressed in Hair Follicles. *J. Invest. Dermatol.* **136**: 1097-105.

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., et al. (2011): Identification of genomic indels and structural variations using split reads. *BMC Genomics* **12**: 375.

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., et al. (2009): A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol.* **10**: R42.

Zuryn, S., Le Gras, S., Jamet, K. & Jarriault, S. (2010): A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186**: 427-30.

10 Author contributions

Author contributions to the scientific publications as part of this habilitation thesis:

Publication 1

Metzger J, Tonda R, Beltran S, Agueda L, Gut M, Distl O: Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. BMC Genomics. 2014;15:562.

Study design: Metzger, Distl
Carried out the experiments: Metzger, Distl, Tonda, Beltran, Agueda, Gut
Data analysis: Metzger, Distl, Tonda, Beltran, Gut
Interpretation of results: Metzger, Distl
Preparation of the manuscript: Metzger, Distl

Publication 2

Metzger J, Karwath M, Tonda R, Beltran S, Agueda L, Gut M, Gut IG, Distl O: Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics. 2015;16:764.

Study design: Metzger, Distl
Carried out the experiments: Metzger, Distl, Karwath, Tonda, Beltran, Agueda, Gut M, Gut IG
Data analysis: Metzger, Distl, Karwath, Tonda, Beltran, Gut IG
Interpretation of results: Metzger, Distl
Preparation of the manuscript: Metzger, Distl

Publication 3

Schrimpf R, Gottschalk M, Metzger J, Martinsson G, Sieme H, Distl O: Screening of whole genome sequences identified high-impact variants for stallion fertility. BMC Genomics. 2016;17:288.

Study design: Schrimpf, Metzger, Sieme, Distl
Carried out the experiments: Schrimpf, Gottschalk, Martinsson, Metzger, Distl
Data analysis: Schrimpf, Gottschalk, Distl
Interpretation of results: Schrimpf, Gottschalk, Metzger, Sieme, Distl
Preparation of the manuscript: Schrimpf, Distl

Publication 4

Metzger J, Rau J, Conn LB, Lindgren G, Distl O: Synergism of four key regulators determines extremely small body size in horses. BMC Genomics, 2018;19(1):492, doi: 10.1186/s12864-018-4877-5.

Study design: Metzger, Distl
Carried out the experiments: Metzger, Rau, Conn, Lindgren
Data analysis: Metzger, Distl
Interpretation of results: Metzger, Distl, Lindgren
Preparation of the manuscript: Metzger, Distl

Publication 5

Metzger J, Pfahler S, Distl O: Variant detection and runs of homozygosity in next generation sequencing data elucidate the genetic background of Lundehund syndrome. BMC Genomics. 2016;17:535.

Study design: Metzger, Distl
Carried out the experiments: Metzger, Pfahler
Data analysis: Metzger, Distl
Interpretation of results: Metzger, Distl
Preparation of the manuscript: Metzger, Distl

Publication 6

Metzger J, Distl O: A study of Shar-Pei dogs refutes association of the 'meatmouth' duplication near HAS2 with Familial Shar-Pei Fever. Animal Genetics. 2014;45(5):763-764.

Study design: Metzger
Carried out the experiments: Metzger
Data analysis: Metzger, Distl
Interpretation of results: Metzger, Distl
Preparation of the manuscript: Metzger, Distl

Publication 7

Metzger J, Nolte A, Uhde AK, Hewicker-Trautwein M, Distl O: Whole genome sequencing identifies missense mutation in MTBP in Shar-Pei affected with Autoinflammatory Disease (SPAID). BMC Genomics. 2017;18(1):348.

Study design: Metzger, Distl
Carried out the experiments: Metzger, Nolte, Uhde, Hewicker-Trautwein
Data analysis: Metzger, Distl
Interpretation of results: Metzger, Nolte, Hewicker-Trautwein, Distl
Preparation of the manuscript: Metzger, Distl

Publication 8

Metzger J, Gast AC, Schrimpf R, Rau J, Eikelberg D, Beineke A, Hellige M, Distl O: Whole-genome sequencing reveals a potential causal mutation for dwarfism in the Miniature Shetland pony. Mammalian Genome. 2017;28(3-4):143-151.

Study design: Metzger
Carried out the experiments: Metzger, Gast, Schrimpf, Rau, Eikelberg, Beineke, Hellige
Data analysis: Metzger, Distl
Interpretation of results: Metzger, Beineke, Hellige, Distl
Preparation of the manuscript: Metzger

Publication 9

Thomer A, Gottschalk M, Christmann A, Naccache F, Jung K, Hewicker-Trautwein M, Distl O, Metzger J: An epistatic effect of KRT25 on SP6 is involved in curly coat in horses. Scientific Reports. 2018;8:6374, doi: 10.1038/s41598-018-24865-3.

Study design: Metzger, Distl
Carried out the experiments: Thomer, Gottschalk, Christmann, Naccache, Hewicker-Trautwein, Metzger
Data analysis: Thomer, Jung, Distl, Metzger
Interpretation of results: Thomer, Distl, Metzger
Preparation of the manuscript: Metzger, Thomer, Distl

Publication 10

Braun M, Reinartz S, Heppelmann M, Rehage J, Sürie C, Distl O, Metzger J: Curly coat caused by a keratin 27 variant was transmitted from Fleckvieh into German Angus. Animal Genetics. 2018, doi: 10.1111/age.12669

Study design: Metzger, Distl

Carried out the experiments: Braun, Reinartz, Heppelmann, Rehage, Sürie, Metzger

Data analysis: Braun, Distl, Metzger

Interpretation of results: Braun, Distl, Metzger

Preparation of the manuscript: Braun, Distl, Metzger

11 Acknowledgements

First and foremost, I want to thank Prof. Dr. Dr. habil. O. Distl for his support during my scientific career and my work at the Institute of Animal Breeding and Genetics. His support, enthusiasm and guidance during all this time were invaluable.

I would also like to thank the German Research Foundation, DFG (ME 4517/1-1) for the financial support of the project “Molecular genetic characterization of curly hair in Curly Horses using marker sets and next generation sequencing data”.

I thank all my collaboration partners for their support in various clinical, pathological, genetic questions and help in lab experiments and data analysis.

I further thank H. Klippert-Hasberg, M. Drabert and N. Wagner for their support in laboratory work and J. Wrede for his tireless help in statistical analyses. I also wish to thank all my colleagues at the Institute of Animal Breeding and Genetics at the University of Veterinary Medicine Hannover who offered advice and support.

I wish to express my particular gratitude to Oliver for his loving support, understanding and for always being there when I needed him. I’m grateful for your love and the time you share with me.

Very special thanks go to my parents for their boundless love and for always standing on my side.

12 Appendix

12.1 Publication 1

Metzger J, Tonda R, Beltran S, Agueda L, Gut M, Distl O: Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. BMC Genomics. 2014;15:562.

Abstract
Background: Domestication has shaped the horse and lead to a group of many different types. Some have been under strong human selection while others developed in close relationship with nature. The aim of our study was to perform next generation sequencing of breed and non-breed horses to provide an insight into genetic influences on selective forces. Results: Whole genome sequencing of five horses of four different populations revealed 10,193,421 single nucleotide polymorphisms (SNPs) and 1,361,948 insertion/deletion polymorphisms (indels). In comparison to horse variant databases and previous reports, we were able to identify 3,394,883 novel SNPs and 868,525 novel indels. We analyzed the distribution of individual variants and found significant enrichment of private mutations in coding regions of genes involved in primary metabolic processes, anatomical structures, morphogenesis and cellular components in non-breed horses and in contrast to that private mutations in genes affecting cell communication, lipid metabolic process, neurological system process, muscle contraction, ion transport, developmental processes of the nervous system and ectoderm in breed horses. Conclusions: Our next generation sequencing data constitute an important first step for the characterization of non-breed in comparison to breed horses and provide a large number of novel variants for future analyses. Functional annotations suggest specific variants that could play a role for the characterization of breed or non-breed horses.

12.2 Publication 2

Metzger J, Karwath M, Tonda R, Beltran S, Agueda L, Gut M, Gut IG, Distl O: Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics. 2015;16:764.

Abstract
Background: Modern horses represent heterogeneous populations specifically selected for appearance and performance. Genomic regions under high selective pressure show characteristic runs of homozygosity (ROH) which represent a low genetic diversity. This study aims at detecting the number and functional distribution of ROHs in different horse populations using next generation sequencing data. Methods: Next generation sequencing was performed for two Sorraia, one Dülmen Horse, one Arabian, one Saxon-Thuringian Heavy Warmblood, one Thoroughbred and four Hanoverian. After quality control reads were mapped to the reference genome EquCab2.70. ROH detection was performed using PLINK, version 1.07 for a trimmed dataset with 11,325,777 SNPs and a mean read depth of 12. Stretches with homozygous genotypes of >40 kb as well as >400 kb were defined as ROHs. SNPs within consensus ROHs were tested for neutrality. Functional classification was done for genes annotated within ROHs using PANTHER gene list analysis and functional variants were tested for their distribution among breed or non-breed groups. Results: ROH detection was performed using whole genome sequences of ten horses of six populations representing various breed types and non-breed horses. In total, an average number of 3492 ROHs were detected in windows of a minimum of 50 consecutive homozygous SNPs and an average number of 292 ROHs in windows of 500 consecutive homozygous SNPs. Functional analyses of private ROHs in each horse revealed a high frequency of genes affecting cellular, metabolic, developmental, immune system and reproduction processes. In non-breed horses, 198 ROHs in 50-SNP windows and seven ROHs in 500-SNP windows showed an enrichment of genes involved in reproduction, embryonic development, energy metabolism, muscle and cardiac development whereas all seven breed horses revealed only three common ROHs in 50-SNP windows harboring the fertility-related gene YES1. In the Hanoverian, a total of 18 private ROHs could be shown to be located in the region of genes potentially involved in neurologic control, signaling, glycogen balance and reproduction. Comparative analysis of homozygous stretches common in all ten horses displayed three ROHs which were all located in the region of KITLG, the ligand of KIT known to be involved in melanogenesis, haematopoiesis and gametogenesis. Conclusions: The results of this study give a comprehensive insight into the frequency and number of ROHs in various horses and their potential influence on population diversity and selection pressures.

Comparisons of breed and non-breed horses suggest a significant artificial as well as natural selection pressure on reproduction performance in all types of horse populations.

12.3 Publication 3

Schrimpf R, Gottschalk M, Metzger J, Martinsson G, Sieme H, Distl O: Screening of whole genome sequences identified high-impact variants for stallion fertility. BMC Genomics. 2016;17:288.

Abstract

Background: Stallion fertility is an economically important trait due to the increase of artificial insemination in horses. The availability of whole genome sequence data facilitates identification of rare high-impact variants contributing to stallion fertility. The aim of our study was to genotype rare high-impact variants retrieved from next-generation sequencing (NGS)-data of 11 horses in order to unravel harmful genetic variants in large samples of stallions. Methods: Gene ontology (GO) terms and search results from public databases were used to obtain a comprehensive list of human und mice genes predicted to participate in the regulation of male reproduction. The corresponding equine orthologous genes were searched in whole genome sequence data of seven stallions and four mares and filtered for high-impact genetic variants using SnpEFF, SIFT and Polyphen 2 software. All genetic variants with the missing homozygous mutant genotype were genotyped on 337 fertile stallions of 19 breeds using KASP genotyping assays or PCR-RFLP. Mixed linear model analysis was employed for an association analysis with de-regressed estimated breeding values of the paternal component of the pregnancy rate per estrus (EBV-PAT). Results: We screened next generation sequenced data of whole genomes from 11 horses for equine genetic variants in 1194 human and mice genes involved in male fertility and linked through common gene ontology (GO) with male reproductive processes. Variants were filtered for high-impact on protein structure and validated through SIFT and Polyphen 2. Only those genetic variants were followed up when the homozygote mutant genotype was missing in the detection sample comprising 11 horses. After this filtering process, 17 single nucleotide polymorphism (SNPs) were left. These SNPs were genotyped in 337 fertile stallions of 19 breeds using KASP genotyping assays or PCR-RFLP. An association analysis in 216 Hanoverian stallions revealed a significant association of the splice-site disruption variant g.37455302G>A in NOTCH1 with the de-regressed estimated breeding values of the paternal component of the pregnancy rate per estrus (EBV-PAT). For 9 high-impact variants within the genes CFTR, OVGP1, FBXO43, TSSK6, PKD1, FOXP1, TCP11, SPATA31E1 and NOTCH1 (g.37453246G>C) absence of the homozygous mutant genotype in the validation sample of all 337 fertile stallions was obvious. Therefore, these variants were considered as potentially deleterious factors for stallion fertility. Conclusions: In conclusion, this study revealed 17 genetic variants with a predicted high damaging effect on

protein structure and missing homozygous mutant genotype. The g.37455302G>A NOTCH1 variant was identified as a significant stallion fertility locus in Hanoverian stallions and further 9 candidate fertility loci with missing homozygous mutant genotypes were validated in a panel including 19 horse breeds. To our knowledge, this is the first study in horses using next generation sequencing data to uncover strong candidate factors for stallion fertility.

12.4 Publication 4

Metzger J, Rau J, Conn LB, Lindgren G, Distl O: Synergism of four key regulators determines extremely small body size in horses. BMC Genomics, 2018;19(1):492, doi: 10.1186/s12864-018-4877-5.

Abstract

Background: Miniature size in horses represents an extreme reduction of withers height that originated after domestication. In some breeds, it is a highly desired trait representing a breed- or subtype-specific feature. The genomic changes that emerged due to strong-targeted selection towards this distinct type remain unclear. Results: Comparisons of whole-genome sequencing data from two Miniature Shetland ponies and one standard-sized Shetland pony, performed to elucidate genetic determinants for miniature size, revealed four synergistic variants, limiting withers height to 34.25 in. (87 cm). Runs of homozygosity regions were detected spanning these four variants in both the Miniature Shetland ponies and the standard-sized Shetland pony. They were shown to be characteristic of the Shetland pony breed, resulting in a miniature type under specific genotypic combinations. These four genetic variants explained 72% of the size variation among Shetland ponies and related breeds. The length of the homozygous regions indicate that they arose over 1000 years ago. In addition, a copy number variant was identified in DIAPH3 harboring a loss exclusively in ponies and donkeys and thus representing a potential height-associated variant. Conclusion: This study reveals main drivers for miniature size in horses identified in whole genome data and thus provides relevant candidate genes for extremely short stature in mammals.

12.5 Publication 5

Metzger J, Pfahler S, Distl O: Variant detection and runs of homozygosity in next generation sequencing data elucidate the genetic background of Lundehund syndrome. BMC Genomics. 2016;17:535.

Abstract

Background: The Lundehund is a highly specialized breed characterized by a unique flexibility of the joints and polydactyly in all four limbs. The extremely small population size and high inbreeding has promoted a high frequency of diseased dogs affected by the Lundehund syndrome (LS), a severe gastro-enteropathic disease. Results: Comprehensive analysis of bead chip and whole-genome sequencing data for LS in the Lundehund resulted in a genome-wide association signal on CFA 34 and LS-specific runs of homozygosity (ROH) in this region. Filtering analysis for variants with predicted high or moderate effects revealed a missense mutation in *LEPREL1* 1.2 Mb proximal to the region of the genome-wide association, which was shown to be significantly associated with LS. LS-affected Lundehund harbored the mutant *LEPREL1*:g.139212C>G genotype A/A whereas all controls of other breeds showed the C/C wild type. In addition, ROH analysis for the Lundehund indicated a high enrichment of genes in potential signatures of selection affecting protein activation and immunoregulatory processes like *NOD1* potentially involved in LS breed disposition. Conclusions: Sequencing results for Lundehund specific traits reveal a potential causative mutation for LS in the neuropeptide operating gene *LEPREL1* and suggests it as a precursor of the inflammatory process. Analyses of ROH regions give an insight into the genetic background of characteristic traits in the Lundehund that remain to be elucidated in the future.

12.6 Publication 6

Metzger J, Distl O: A study of Shar-Pei dogs refutes association of the 'meatmouth' duplication near HAS2 with Familial Shar-Pei Fever. Animal Genetics. 2014;45(5):763-764.

Abstract

No abstract available.

12.7 Publication 7

Metzger J, Nolte A, Uhde AK, Hewicker-Trautwein M, Distl O: Whole genome sequencing identifies missense mutation in MTBP in Shar-Pei affected with Autoinflammatory Disease (SPAID). BMC Genomics. 2017;18(1):348.

Abstract

Background: Autoinflammatory diseases in dogs are characterized by complex disease processes with varying clinical signs. In Shar-Pei, signs of inflammation including fever and arthritis are known to be related with a breedspecific predisposition for Shar-Pei Autoinflammatory Disease (SPAID). Results: Clinical and histopathological examinations of two severely SPAID-affected Shar-Pei revealed signs of inflammation including fever, arthritis, and perivascular and diffuse dermatitis in both dogs. A multifocal accumulation of amyloid in different organs was found in one SPAID-affected case. Whole genome sequencing resulted in 37 variants, which were homozygous mutant private mutations in SPAID-affected Shar-Pei. Nine SNVs with predicted damaging effects and three INDELs were further investigated in 102 Shar-Pei affected with SPAID, 62 unaffected Shar-Pei and 162 controls from 11 different dog breeds. The results showed the missense variant MTBP:g.19383758G > A in MTBP to be highly associated with SPAID in Shar-Pei. In the region of this gene a large ROH (runs of homozygosity) region could be detected exclusively in the two investigated SPAID-affected Shar-Pei compared to control dog breeds. No further SPAID-associated variant with predicted high or moderate effects could be found in genes identified in ROH regions. This MTBP variant was predicted to affect the MDN2-binding protein domain and consequently promote proinflammatory reactions. In the investigated group of Shar-Pei older than six years all dogs with the mutant genotype A/A were SPAID-affected whereas SPAID-unaffected dogs harbored the homozygous wildtype (G/G). Shar-Pei with a heterozygous genotype (G/A) were shown to have a 2.13-fold higher risk for disease development, which gave evidence for an incomplete dominant mode of inheritance. Conclusions: The results of this study give strong evidence for a variant in MTBP related with proinflammatory processes via MTBP-MDM2 pathway. Thus, these results enable a reliable detection of SPAID in Shar-Pei dogs.

12.8 Publication 8

Metzger J, Gast AC, Schrimpf R, Rau J, Eikelberg D, Beineke A, Hellige M, Distl O: Whole-genome sequencing reveals a potential causal mutation for dwarfism in the Miniature Shetland pony. Mammalian Genome. 2017;28(3-4):143-151.

Abstract

The Miniature Shetland pony represents a horse breed with an extremely small body size. Clinical examination of a dwarf Miniature Shetland pony revealed a lowered size at the withers, malformed skull and brachygnathia superior. Computed tomography (CT) showed a shortened maxilla and a cleft of the hard and soft palate which protruded into the nasal passage leading to breathing difficulties. Pathological examination conirmed these findings but did not reveal histopathological signs of premature ossification in limbs or cranial sutures. Whole-genome sequencing of this dwarf Miniature Shetland pony and comparative sequence analysis using 26 reference equids from NCBI Sequence Read Archive revealed three probably damaging missense variants which could be exclusively found in the afected foal. Validation of these three missense mutations in 159 control horses from diferent horse breeds and five donkeys revealed only the aggrecan (ACAN)-associated g.94370258G>C variant as homozygous wild-type in all control samples. The dwarf Miniature Shetland pony had the homozygous mutant genotype C/C of the ACAN:g.94370258G>C variant and the normal parents were heterozygous G/C. An unafected full sib and 3/5 unaffected half-sibs were heterozygous G/C for the ACAN:g.94370258G>C variant. In summary, we could demonstrate a dwarf phenotype in a miniature pony breed perfectly associated with a missense mutation within the ACAN gene.

12.9 Publication 9

Thomer A, Gottschalk M, Christmann A, Naccache F, Jung K, Hewicker-Trautwein M, Distl O, Metzger J: An epistatic effect of KRT25 on SP6 is involved in curly coat in horses. Scientific Reports. 2018;8:6374, doi: 10.1038/s41598-018-24865-3.

Abstract

Curly coat represents an extraordinary type of coat in horses, particularly seen in American Bashkir Curly Horses and Missouri Foxtrotters. In some horses with curly coat, a hypotrichosis of variable extent was observed, making the phenotype appear more complex. In our study, we aimed at investigating the genetic background of curly coat with and without hypotrichosis using high density bead chip genotype and next generation sequencing data. Genome-wide association analysis detected significant signals ($p = 1.412 \times 10^{-05}$–$1.102 \times 10^{-08}$) on horse chromosome 11 at 22–35 Mb. In this significantly associated region, six missense variants were filtered out from whole-genome sequencing data of three curly coated horses of which two variants within KRT25 and SP6 could explain all hair phenotypes. Horses heterozygous or homozygous only for KRT25 variant showed curly coat and hypotrichosis, whereas horses with SP6 variant only, exhibited curly coat without hypotrichosis. Horses with mutant alleles in both variants developed curly hair and hypotrichosis. Thus, mutant KRT25 allele is masking SP6 allele effect, indicative for epistasis of KRT25 variant over SP6 variant. In summary, genetic variants in two different genes, KRT25 and SP6, are responsible for curly hair. All horses with KRT25 variant are additionally hypotrichotic due to the KRT25 epistatic effect on SP6.

12.10 Publication 10

Braun M, Reinartz S, Heppelmann M, Rehage J, Sürie C, Distl O, Metzger J: Curly coat caused by a keratin 27 variant was transmitted from Fleckvieh into German Angus. Animal Genetics. 2018, doi: 10.1111/age.12669

Abstract
No abstract available.