BIOINFORMATICS' APPROACHES TO DETECT GENETIC VARIATION IN WHOLE GENOME SEQUENCING DATA

Thesis Committee

Thesis supervisors

Prof. dr. Martien A.M. Groenen
Personal chair at the Animal Breeding and Genomics Centre
Wageningen University

Prof. dr. Mari A. Smits
Personal chair at the Animal Breeding and Genomics Centre
Wageningen University

Other Members

Dr. Roeland van Ham (Wageningen University, the Netherlands)
Prof. dr. Jack A.M. Leunissen (Wageningen University, the Netherlands)
Prof. dr. Johan T. den Dunnen (Leiden University Medical Center, the Netherlands)
Prof. dr. Edwin Cuppen (Hubrecht Institute, Utrecht, the Netherlands)

This research was conducted under the auspices of the Graduate school of Wageningen Institute of Animal Sciences (WIAS).

Bioinformatics' approaches to detect genetic variation in whole genome sequencing data

Hindrik H.D. Kerstens

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. dr. M.J. Kropff
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday October 25th 2010
at 1.30 p.m. in the Aula.

Hindrik Harm Dirk Kerstens (2010)
Bioinformatics' approaches to detect genetic variation in whole genome sequencing data
Thesis Wageningen University, Wageningen, the Netherlands
With references, with summary in English and Dutch

ISBN: 978-90-8585-780-8

Contents

Chapter 1 Introduction.....7
Chapter 2 Mining for single nucleotide polymorphisms in pig genome sequence data.....43
Chapter 3 Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey.....61
Chapter 4 Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries.....85
Chapter 5 Genome wide SNP discovery and analysis in mallard (Anas platyrhynchos).....115
Chapter 6 Discussion.....137
Summary.....163
Samenvatting.....167
Dankwoord.....173
About the Author.....177
List of Publications.....179
Training and Supervision Plan.....183
Colophon.....185

1 Introduction

Introduction

The description of the DNA structure in 1953 marked the beginning of the subsequent identification of genes and their functions. Over the past decades this information enabled the detection of genetic differences among individuals on the DNA level underlying many important traits. In farm animal breeding, the benefit of using genetic differences on the DNA, that are linked to genes of interest, is the possibility to select for distinguishable genetic characteristics (genotypes) instead of the selection based on (visible) characteristics (phenotypes). How the availability of a genetic marker map can contribute in developing a research tool facilitating discovery of genetic factors that contribute to susceptibility to disease, to protection against illness or to drug response, is illustrated in human by means of the haplotype map of the human genome (HapMap) project [1].The availability of large numbers of common DNA sequence variants in the human genome made it possible to use this variation for genome wide association (GWA) studies. This enabled the establishment of clear associations between particular genomic regions and a phenotype of interest, for example for diseases like diabetes type 2 and Myocardial Infarction [2,3]. Furthermore, the publicly available HapMap data offers further opportunities in e.g. customization of medical treatments based on a patient's genetic make-up, thereby maximizing the effectiveness of the therapy and minimizing side effects for the patient.

Similar initiatives are currently deployed in a number of farm animals, the bovine HapMap Consortium for example already published a map of the genetic diversity among different bovine populations [4]. Detailed analyses of the genetic variation in farm animal populations, enable further insight in the genetic diversity of ancestral population and the effects of domestication, and selection, on this genetic diversity. With the availability of sequence variation data, also important traits like reproduction and resistance to livestock diseases can be studied and the ongoing development of DNA techniques allow for improved strategies to meet the breeding goals. The development of large genotyping assays and automated genotyping tools allow for detailed genetic screening and can help to further optimize population management methods for the selection of new or improved traits. Currently animals can be selected based on their predicted genetic value using genetic markers found throughout the genome. This approach is made possible by the availability of gene chips with tens of thousands of features capable of measuring single nucleotide polymorphism (SNP) markers for a growing number of species. However for

most farm animals the number of SNPs identified is rather limited (Table 1). This thesis addresses the challenge of a reliable, high throughput and cost-effective identification of genetic markers in DNA sequencing data. We developed and validated strategies for detecting SNPs in genomes of animals using traditional and current second generation sequencing data. Developed SNP resources may be applied for the identification of trait loci and the generation of linkage maps. These SNP recourses have been made publicly available and are a big step forward towards fulfilling the requirements for performing GWA studies in the animal species we analyzed. We also provided a first glimpse in abundance of structural variants (SVs), a marker so far not commonly used in animal genetics, which can be detected by an alternative use of data resulting from current second generation DNA sequencing techniques. In the next paragraphs an overview will be given of applications in animal breeding in which genetic markers are being used and how markers are obtained.

Table 1: Exponentially growing number of SNPs in dbSNP.

	Number of SNP Clusters	
Organism	***April 2008***	***March 2010***
Human	14,708,752	105,098,087
Chicken	3,293,383	11,318,091
Cow	2,223,033	4,931,121
Pig	8,427	542,119
Turkey	0	8,536[1]
Duck	0	0[2]

The increase in the number of SNPs in pig include 6374 SNPs identified by our study (chapter 2 of this thesis).1=Turkey SNPs (chapter 3 of this thesis) are scheduled to be included for next (132) dbSNP release . 2=Duck SNPs (chapter 5 of this thesis) have not been submitted yet.

Markers in animal breeding

For centuries human are exploiting variation in plant and animal genetic resources important for producing food and other agricultural products. Through breeding, human have succeeded to significantly improve and secure production of these products. Initially, breeding progress was achieved by

selecting and reproducing preferred individuals based on phenotypic records of the individual and its relatives. The application of Mendel's laws of inheritance and the inclusion of crosses in breeding schemes improved the selection for desired traits. The continued progress made in research and technology has allowed the production of improved varieties and breeds with increasing accuracy and efficiency. However, characteristics influenced by the environment, multiple genetic factors or with low heritability, are difficult breeding targets. Furthermore traits that are difficult to measure, that are only visible late in life or in only one of the sexes, and traits that require sacrifice or challenge of the animal, are costly to screen for using phenotype information only. Examples of these traits are health and welfare traits like resistance to mastitis in dairy cows and traits related to sustainability like livestock fertility and intestine function.

The discovery of DNA and the established relationship between the information contained in DNA and the structure of proteins, which determine the characteristics of an individual allowed for a shift from phenotype-based towards genotype-based selection. Differences in the DNA information, some of which causing phenotypic differences, are named genetic differences or molecular markers. Molecular markers are DNA sequence variations, found at specific locations in the genome, and inherited by the Mendelian laws. They can be seen as landmarks on the genome that can be identified by molecular assays. Techniques to identify tens to hundreds of markers were for the first time described in the late 1970s [5] allowing the use of markers to detect associations with traits of interest [6].

Marker Assisted Selection

The idea behind marker assisted selection is that specific genes have a significant effect on specific traits and that these can be targeted specifically during selection procedures. Using a marker map, putative genes significantly affecting traits of interest can be detected by testing for statistical associations between marker variants and a trait of interest. The availability of at least a sparse map of genetic differences (markers) together with information on their association with the animals' phenotype is used in marker assisted selection (MAS). In MAS, traits of interest are selected indirectly by not only selecting the trait itself (by phenotype) but also on a marker genetically linked to it.

However, relatively few traits in farm animals are controlled by a single or only a few genes as has been found for hair/plumage color in animals [7-10] or glycogen content in muscle [11]. The majority of traits of economic importance

in farm animals are genetically complex quantitative traits that most likely are controlled by a large number of genes and regulatory elements. A chromosomal region that contains one or more genes or regulatory elements that influences a (multifactorial) trait is known as a quantitative trait locus (QTL)[12]. The identification of QTL requires statistical analysis to demonstrate that specific genes or genomic regions have a significant effect on the phenotype. Typically QTL studies result in multiple regions on the genome each containing one or multiple genes that are associated with the trait being assayed or measured.
QTL detection can theoretically be applied to any species for which a molecular marker map is available. For many farm animals including chicken, cattle, pig and sheep molecular marker maps are available and many QTL have been identified [13-16]. Identified QTL are publicly available in the animal QTL database (http://www.animalgenome.org/QTLdb/).[17]. However, due to the limited number of markers, an thus the limited resolution, the genomic locations of QTL can only be given with rather large confidence intervals. Recent advances in molecular genetic techniques have made dense marker maps available and have made genotyping of many individuals for these markers feasible. These developments had only a minor effect on the precision of mapping QTL by traditional linkage analysis [18]. Furthermore the identified QTL explain only a limited fraction of the genetic variation of the trait of interest. Within a QTL, some of the genes can have a relatively large effect and are referred to as major genes located at QTL. Effectively, only genes with large effects will be detected and mapped by QTL mapping although the term QTL strictly applies to loci of any effect. Therefore, a different approach named genome wide marker assisted selection (GWMAS) is needed to perform selection based on marker genotypes by optimally using all marker information.

GWMAS

In genome wide marker assisted selection (GWMAS) the significance testing is omitted by simply estimating the effects of all genes or chromosomal positions simultaneously [19]. Required dense marker maps provide many markers in and surrounding QTL regions which are all included in genome wide selection. In other words genome wide selection is a marker-based selection without first identifying a subset of markers with significant effects. Initially genotype-phenotype relationships are established in reference populations. By using all the obtained marker-phenotype information, additive genetic effects of markers, also called breeding values, of individuals without phenotypic record and no progeny can be estimated. Subsequently, selection based on these breeding

values can be applied to substantially increase the rate of genetic gain in animals and plants, especially if combined with reproductive techniques to shorten the generation interval. The genome wide selection approach has proven to be useful in dairy cattle breeding [20] and in plants for example it outperformed Marker Assisted Recurrent Selection in Maize [21].

The role of markers in genome wide association studies

The aim of GWA studies is to find associations between the genotypes at each locus and presence of a trait. GWA studies in humans have already successfully identified a number of markers in or close to genes that are involved in common human diseases [22-24] and have provided insights on novel pathways and potential therapeutic targets [25]. In animal breeding large scale GWA studies have been applied in cattle to identify genomic locations associated with growth [26].

Genome sequencing, DNA marker identification and large scale genotyping led to the discovery that human and animal genomes consist of an alternating series of blocks of high linkage disequilibrium (LD) [4,27-30] and recombination hotspots [31,32].Due to the existence of LD (a non-random association of alleles at two or more loci) common variation in the genome can be surveyed by genotyping only a fraction of the total number of genomic variants that exist in the population [33]. An association study using indirect approach is able to capture most sequence variation because insights from human population genetics suggest that ~90% of sequence variation among individuals is due to common variants [34]. In cattle such an approach is also feasible since the cattle and human genomes show similar LD and haplotype block structure [35]. Due to a relatively low mutation rate, most of the common variants originally arose from single historical mutation events. Common variants are therefore associated with neighboring variants that were already present on the ancestral chromosome. These associations allow the identification of candidate genes by using information from a relatively small set of variants representing most of the common patterns of variation in the genome. This provides an important shortcut to carry out candidate-gene and GWA studies in a certain population by minimizing the numbers of SNPs that need to be genotyped [36].

The aim of the International HapMap Projects is to determine the common patterns of DNA sequence variation in the human and animal genomes, by characterizing sequence variants, their frequencies, and correlations between them, in DNA samples from populations with ancestry. Haplotypes, which are defined as a set of SNPs on a single chromatid that are statistically associated,

are identified in human [36] and cattle [4]. HapMap projects for pig and chicken are currently in progress.
The resulting haplotype maps provides a framework for studying associations between genes and phenotypes. By making use of LD, genotyping demands are reduced while still providing an equivalent power as threefold more randomly chosen markers [37]. Haplotype blocks, associated with a certain phenotype, can subsequently be followed over generations and will gradually decay through each generation due to recombination, but linkage will persist for closely linked loci. SNPs and sequence analysis are then used to define the minimum haplotype that is shared identical by descent among animals showing a certain phenotype [38].

Marker based comparative genomics in studying genome evolution

Comparative genomics allows the transfer of genome information from a well-characterized species to another genome that is less well characterized. It can be applied at all genomic levels, from chromosome maps to the completely sequenced genome. In comparative genomics sequenced genomes are compared and the information provided by genome evolution is used to better understand genome function and evolutionary processes that act on genomes. Despite the recent advances in sequencing technologies, the remaining challenges in producing a genome sequence assembly, in particular from short sequencing reads, is reflected by the small number of vertebrate genomes whose genome has been sequenced to date. In birds, for years there was only a single published genome sequence, that of the chicken [39] recently accompanied by the zebra finch genome [40]. The completion of the chicken genome and its associated resources like mapped BACs, markers, linkage map and genome annotation facilitates the rapid development of detailed genomic information, potentially in all other birds. Powerful strategies are combinations of in-silico and in-vitro approaches involving sequence comparison, cross-species fluorescent in-situ hybridization (zoo-FISH) [41,42] and the use of whole-genome tiling path micro-arrays for cross-species array comparative genomic hybridization (aCGH) [43-45]. Such a combination of techniques provides information on gross genomic rearrangements, gene gains/losses, copy number variation (CNV) and gene order conservation. These techniques do not require sequence data for any species other than the reference (i.e. chicken, zebrafish or human) and have previously successfully been applied for a genome wide comparison of chromosomal rearrangements and CNVs between e.g. chicken and Peking

duck [43] chicken and turkey [46], zebrafish and catfish [47] and human and primates [48]. Results revealed a strong conservation of genome structure over tens of million years of evolution between compared species. Comparative genomics therefore accelerates mapping studies in the studied species not only by facilitating the transfer of genetic information in the form of markers but also gene predictions directly from the closest sequenced reference species to the studied species. Genetic maps, which document the way in which recombination rates vary over a genome, are also an essential tool in studying genome evolution. Markers can be used to construct a haplotype map facilitating the study of genome evolution by means of recombination events. The required marker density depends on LD within the studied population which is determined by the effective population size, mutation rate and recombination rate. The latter varies largely within species: e.g. ~0.4 cM Mb-1 in mouse, ~0.8 cM Mb-1 in human and ~2 cM Mb-1 within chicken [49]. Moreover within a genome recombination rates for chromosomes can differ up to an order of magnitude with a relative excess of recombination on smaller chromosomes.

SNPs, a valuable and efficient molecular marker

Single nucleotide polymorphisms (SNPs) are increasingly finding their application in studies of genetic variation within populations. More than a decade ago, SNP detection was laborious and genotyping cost were generally to high (more than US$1 to obtain one data point) to perform large scale genotyping studies using SNPs. The introduction of high-throughput molecular marker discovery and the advances in high-throughput genotyping have reduced the cost per data point considerably. This reduction was mainly the result of three parallel developments [50,51]: (i) the discovery of vast numbers of single nucleotide polymorphism (SNP) markers in many species; (ii) development of high-throughput technologies, such as multiplexing and gel-free DNA arrays, for screening SNP polymorphisms; and (iii) automation of the marker-genotyping process, including streamlined procedures for DNA extraction. These developments have recently made the SNP marker more popular than other kinds of existing molecular markers, such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs) and microsatellites. The major advantages of using SNPs are their technical requirements (they can be automated and do not require use of radioactivity) and the relatively small investment of time, money and labour to develop SNP assays and perform

genotyping. Probably, the most important feature is their occurrence varying from on average one in 1300 bp in human [52], one in 300 to 400 bp in pig [53] and one in 200 bp in chicken [39] indicating that there are millions of SNPs on a single genome. SNPs are DNA single nucleotide (A,T,C or G) variations that occur, between alleles of an individual or between individuals. To be considered as a SNP, by definition such variation must occur in at least 1% of the population. Although the majority of the SNPs have no known biological effect, some variations in DNA sequence are related to disease or cause a specific phenotype. SNPs are evolutionarily stable, making them suitable markers in population studies.
The availability of cheap and abundant molecular markers like SNPs has changed the use of molecular markers in farm animal breeding programs. Once identified, millions of SNPs within a genome can be used to facilitate mapping of complex traits because all genomic regions involved are tagged by at least one SNP marker and thus can be traced in the pedigree. SNPs are no longer only an aid to the breeding process, but breeding strategies are adjusted to fully exploit SNPs to optimize breeding progress.

Techniques, challenges and threats in obtaining SNPs

SNPs need to be identified before they can be used in genotyping assays. In contrast to human, the number of SNPs in the public databases is still very limited for most farm animals (Table1). Furthermore, at the start of this research project, the chicken genome was the only available sequenced reference genome of a farm animal species. Nevertheless, for some farm animals, genomic sequence reads originating from different individuals and representing the same genomic region were publicly available and these are a potential source to identify genetic variation in these species. For these species a SNP resource can be developed by applying computational methods on pre-existing sequencing and mapping data. In other species genome sequencing has not even started, requiring the development of additional cost-effective sequencing, mapping and subsequent SNP detection approaches.

Mining SNPs from DNA Sequence Data Computationally

The alignment of multiple sequence reads representing the same region on the genome enables the identification of single nucleotide polymorphisms [54]. Where available it is wise to gain access to the primary sequencing data from which base-calling scores automatically can be derived [55] and included these in the analysis. Base-calling scores provide additional information to determine

the probability that identified differences represent true polymorphisms [56]. Observed sequence variants with high-quality scores are likely to represent true polymorphisms whereas low-quality variants have a high chance of being sequencing errors. When retrieving sequence data from public databases, quality scores are not always available. Increasing the sequence coverage such that putative polymorphisms are confirmed by overlapping reads can compensate for this lack of information, since the chance that two independent sequence reads have a sequencing error at the same genomic position is small.
Commonly used software tools for multiple sequence alignment of reads are Phrap [57] and CAP3 [58]. These programs are able to reconstruct contiguous stretches of DNA sequence from smaller (partly) overlapping reads. The presence of single nucleotide substitutions can automatically be detected by softwares like PolyPhred [59] and Polybayes [60]. A more recent version of PolyPhred [61] and other softwares including SNPdetector [62], novoSNP [63] and PolyScan [64] have been developed to detect heterozygous polymorphisms in chromatogram files as a result of direct sequencing of PCR-amplified sequences from diploid samples. Because of their entirely automated fashion, base calling (Phred), assembly (Phrap) and single nucleotide substitution detection (Polybayes) can be put in a de novo sequence based SNP detection pipeline [65] to reliably detect SNPs from clustered EST sequences without manual intervention. As stated earlier, a lack of base-calling information can be compensated by requiring a SNP redundancy score. Additionally the co segregation of the candidate SNP with other SNPs in the alignment is a measure of confidence [66]. A SNP detection strategy in which haplotypes were mathematically reconstructed before the actual SNP calling was performed further reduced the number of false positive SNP predictions [67]. Other SNP detection pipelines like SsahaSNP (Ning unpublished) require, in contrast with de novo sequence based SNP detection approaches, a reference genome. Clusters of homolog sequences, are built by mapping whole genome shotgun reads by using fast search algorithms [68] and sequence polymorphism are consequently identified [69]. Homology search tools like BLAT [70] and Megablast [71] are alternatives to perform this tasks efficiently.

Impact of Next Generation sequencing on SNP discovery

A quite recent source of obtaining affordable sequence data are high throughput next-generation sequencing (NGS) technologies (Illumina GA/Solexa, RocheGS/454, Solid, Helicos). The massively parallel approaches in these technologies allows much faster and more cost-effective sequence

determination than the traditional dideoxy chain terminator method described by Sanger in 1977. The throughput for these new sequencing approaches is measured in billions of base pairs per run, thousand times more than the daily output of an automated 96 capillary DNA sequencer using dye-terminator sequencing technology. High throughput is achieved by immobilization of 400 thousand (Roche) to 40 million (Illumina) target DNAs and sequencing these fragments simultaneously. Furthermore, these new technologies allow the direct sequencing of DNA or cDNA without any cloning step and at decreasing cost per sequenced base (Table 2). Although all of these new technologies produced extremely short reads when they initially entered the market, the performance of (NGS) platforms have increased substantially. As an example: the first next-generation sequencer released by 454 (GS20) had an average read-length of 110 bp. A second improved sequencer has since then been introduced, the GS-FLX, which is able to obtain average read lengths of 250 bases and is able to perform mate-paired reads. A more recent release is the 454 FLX Titanium which has an average read length of 400 bp and it is not unlikely that future releases of NGS platforms will produce read lengths comparable to those produced with traditional dye-terminator sequencing technology.

Table 2: Comparison of sequencing technologies in spring 2008.

Technology	*Approach*	*max througput (bps)*	*read length*	*% accuracy*	*paired-end*	*$/Mbp*
Sanger ABI3730xl[1]	synthesis with dye terminators	1 Mbp/day (12)	800	99.0->99,999	no	1000
454/Roche FLX[2]	pyrosequencing	100 Mb/7.5hr (3.7K)	250	96.0-99.5	2x110	30
Illumina/Solexa[3]	sequencing by synthesis	3 Gbp/5days (6.9K)	36	96.2-99.7	2x36	2.10
ABI/SOLiD[4]	sequencing by ligation	3 Gbp/5days (6.9K)	35	99.0->99.94	2x25	1.30
Helicos[5]	single-molecule sequencing	7.5 Gbp/4days (22K)	25	93->99.0	no	n/a

Throughput is compared by giving the maximum number of bases per second a platform generates at it's optimal readlength. Paired-end sequencing is for some platforms restricted to a read length. Sequencing cost are compared by indicating the price per mega base and only includes the costs of reagents and costs to perform one sequencing run.

[1]Applied Biosystems http://www.appliedbyosystems.com
[2]Roche Applied Science http://www.roche-applied-science.com
[3]Illumina, Inc. http://www.illumina.com
[4]Applied Biosystems http://www.appliedbyosystems.com
[5]Helicosbio http://www.helicosbio.com

NGS technologies allow cost effective sequencing of multiple individuals which is beneficial for SNP detection. SNP discovery through parallel pyrosequencing (454) of an individual human genome identified 3.32 million SNPs, with 606,797 of those as novel SNPs [72] which is comparable with the shotgun-sequenced Venter genome (cost $70 million) that had 3.47 million SNPs, with 647,767 of those being novel [72]. These authors stated that at least a 20X genome coverage is required to call 99% of heterozygous bases correctly within a single individual, resulting in a sequencing cost of 2 million dollar. For the Illumina method similar results (3.07 million SNPs of which 420 thousand novel SNPs) can be obtained by sequencing an individual genome with 36X coverage for less than half a million dollars [73] or ~ 4 million SNPs of which 26% are novel SNPs at 30X coverage for a quarter of a million dollar. [74]. More recent developments using single molecule sequencing further reduce the sequencing costs. For example, the discovery of 2.8 million SNPs by whole genome resequencing at 28X coverage using this technology reduces the costs to less than 50 thousand dollars [75]. These developments indicate that extremely high throughput sequencing machines that produce relatively short reads are favorable for SNP discovery in a whole genome resequencing approach. In addition, simulations suggest that 85% of 35 bp reads can be placed uniquely on the human genome whereas 95% of 35 bp paired end reads with 200 bp insert sizes have unique placement [76]. However SNP discovery in species with limited public genomic resources benefits from longer read lengths. It has been demonstrated that reads produced with pyrosequencing technology can be assembled de novo into reasonably long contigs that subsequently can serve as a genome reference on which short reads of other individuals can be mapped to detect SNPs [77]. At present whole genome sequence assembly by alignment of relative short NGS fragments without the availability of a reference genome is tedious, but possible for less complex mammalian sized genomes [78]. Besides the consistent pattern of non-uniform sequence coverage each NGS platform generates [79], which is substantially lower in AT-rich repetitive sequences, repetitive regions are hard to reconstruct by short-read sequence assembly. Possible NGS SNP detection strategies circumventing the requirement of a sequenced reference genome are (ultra short read) sequencing of more than one genotype and alignment of that data by using: (1) genome or transcriptome sequence data from model species closely related to the species of interest, (2) whole transcriptome or reduced representative genome sequence data for the species of interest, based on Roche/454 sequence technology [80,81].

Systematic reduction of genome complexity towards a cost effective SNP detection approach

While sequencing of a mammalian sized genome remains costly, even using a NGS approach, sequencing at a deep depth on a subset of a genome is certainly in reach and will enable the identification of a considerable amount of genetic variation. Methods dealing exclusively with the expressed fraction of the genome include the sequencing of cDNA, which is DNA reverse transcribed from a mature mRNA, and transcribed short sub-sequences of cDNA that are named Expressed Sequence Tags (ESTs). Methods used for a systematic reduction of genome complexity include technologies such as Amplification Fragment Length Polymorphism (AFLP) [82], Complexity Reduction of Polymorphic Sequences (CroPS) [83] and Reduced Representation Libraries (RRL) [84]. Other methods are targeting a unique fraction of the genome, e.g. Cot Filtration [85]. By using the methods AFLP and RRL a set of relatively short genomic fragments are being sampled whereas Cot filtration results in a set of non repetitive genomic fragments. In particular for highly repetitive genomes the latter method facilitates sequence assembly and variant discovery.

Sequencing errors: a serious threat in SNP discovery

Errors in sequence reads are a serious threat in SNP discovery and subsequent marker development. Any sequencing error can give rise to false positives compromising downstream studies. When sequence information is generated, nucleotide sequence data is accompanied by probability values for each base call. The universally used metric for base calling quality is the “phred score” derived from the base-calling software Phred [55]. The phred score can be calculated by the equation -10log10 P(the base calling is false). The sequence accuracy of a single read is limited by the fidelity of the polymerase used in the amplification and/or readout step of the sequencing technology which is on the order of 10-5-10-7per base pair. In the past decades Sanger style capillary sequencing machines have set the standard for accuracy at a minimum phred score of 20 or 99% accuracy per base. However, the accuracy between read positions 100 and 700 is typically greater then phred score 50, or 99,999%. These figures are in sharp contrast with the accuracy currently obtained by NGS platforms (Table 2). Even for SNP discovery we can deal with this low single read accuracy by using the overall consensus accuracy, which is the composite accuracy from sequencing a specific base at sufficient depth. A sequencing depth of three provides theoretically a 99,9% consensus accuracy at a the single read accuracy of 90%. Therefore NGS platforms that show coverage variability

perform the worst, whereas platforms, with the most uniform coverage, perform the best in sequence accuracy [79]. Errors that persist at high coverage are systematic, platform specific and typically associated with certain sequence contexts [79,86,87]. The combination of coverage uniformity and platform specific biases determine what average target coverage is required to perform variant discovery with an acceptable (low) false positive rate. For an accurate detection of biallelic sites the average depth of sequence coverage required, especially for the short-read technologies, is about 3-5 times higher than the empirically determined coverage of 20-fold utilizing traditional Sanger sequencing [79,88].

Besides sequencing errors also mapping errors, due to short reads and limited sensitivity of less computational intensive alignment algorithms, cause false positive SNP predictions in NGS data. The increasing sequencing throughput forces the replacement of the first generation short read alignment programs MAQ [76], SOAP [89], ELAND (Illumina), by a new generation of short read alignment programs including BOWTIE [90], BWA [91] and SOAP2 [92]. Currently methods are being developed for SNP detection in short read mapping data, carefully considering the data quality, alignment, and experimental errors common to NGS technology [92]. Further improvements in read length and base calling accuracy and the exploitation of paired end information will allow the move to less computational intense algorithms requiring less memory but without the loss of sensitivity.

Structural variants, another source of genetic variation

Structural variants (SV), in particular those larger than one kilobase which are Copy Number Variants (CNVs) are one of the primary genomic mechanisms thought to underlie the evolutionary expansions of species in the past 90 million years. For many years gene duplication is thought to be the driving force in evolutionary change [93,94] a theory for which evidence is accumulating by a growing wealth of (comparative) genomic data. The high abundance of SV in genomes has made scientists curious about the contribution of structural variation to disease traits and adaptation of groups of a species to specific environmental conditions. Structural variation includes insertions, deletions, inversions and translocations in a size range of a few bases to hundreds of kilobases. SV formation occurs by both recombination based and replication based mechanisms and de novo locus specific mutation rates appear much higher for SVs than for SNPs. Because of their high mutation rate which is expected to range from $1.7 \times 10{-}6$ to $1.0 \times 10{-}4$ per locus per generation

[95,96], 100 to 10,000 times higher than nucleotide substitution rates, genomic rearrangements occur frequent enough that both inherited and de novo events can be observed in the same family [97]. However, the contribution (relative to SNP) of SV to locus-specific mutation rate may vary throughout the genome. Analogously to CpG dinucleotides that are hotspots for base substitutions [98,99], flanking low-copy repeats trigger non-allelic homologous recombination and can be considered the 'hotspots' for CNV [95]. In that sense the relative contribution of SV to mutation rate can reflect local genome architecture, resulting in regional susceptibility to genome instability [100]. In the Human genome SVs are common and likely to encompass more polymorphic basepairs ,~0.7% of the genome, than SNPs [101,102].

SVs and phenotypes

Large scale identification [103-106] of the extensive presence of SVs in the human genome and subsequent association studies show that instances of CNVs are related to human health and common genetic diseases like Parkinson's disease, Autism, Psoriases and Rheumatoid Arthritis [107-110]. Phenotypes that have been related to SVs and are caused by genomic rearrangements, can be the consequence of a variety of molecular mechanisms like gene dosage, gene interruption, gene fusion, position effects, unmasking of recessive alleles or functional polymorphism, and potential transvection effects [111].

CNVs involving dosage sensitive genes, such as PMP22, can alter gene expression levels and cause consequent clinical phenotypes such as CMT1A (PMP22 over expression) and HNPP (PMP22 under expression) [112,113]. When the breakpoint of a SV is located within a functional gene, it may interrupt the gene and cause a loss of function as exemplified by the phenotype color blindness [114]. SVs can also result in a biological functional rearrangement between different genes or their regulatory sequences which has been found as a causative mutation in hypertension [115] and is called a gain-of-function mutation. Furthermore a SV can have an effect on expression or regulation of a nearby gene by removing or altering a regulatory sequence which is illustrated in one of the causes of campomelic dysplasia [116] as well as in other human diseases [117]. A SV resulting in a deletion of one allele may unmask another recessive allele or functional polymorphism. In patients with the common Sotos syndrome deletion, for instance, the activity of the plasma coagulation factor 12 is predominantly determined by the functional polymorphism of the remaining hemizygous allele [118]. SV in regulatory elements required for communication between alleles have been shown to mediate in transvection [119], that is the influence on gene expression by the

pairing of alleles on homologous chromosomes [120].

SVs and their role in evolution

In an evolutionary context, SVs can also be potentially exposed to selection pressure during evolution, including purifying and positive selection. In the human genome, the majority of SVs are located outside genes and ultra conserved elements. Furthermore a significantly lower proportion of deletions than duplications overlaps with disease-related genes and RefSeq genes suggesting purifying selection [102,121]. Purifying selection has also been observed in Drosophila, where duplications outnumbered deletions especially in functionally constrained regions [122]. Evidence for positive selection on SVs, in particular gene duplication, has been found by means of human-specific or primate-specific gene amplification. It was found that more than one fourth of the examined human genes represent CNVs in one or more of the 10 primate species and that gene gains typically outnumbered losses, suggesting positive selection in primate genome evolution [45].

In particular copy number differences caused by lineage-specific duplications, potentially altered the expression spectrum of duplicated genes which might have resulted in the acquisition of new functions and therefore of adaptive evolution. The strongest evidence comes from the multiple copy protein domain, DUF1220. The copy number of DUF1220 was shown to be highly expanded in humans, reduced in African great apes, further reduced in orangutan and Old World monkeys, only single-copy in non primate mammals, and absent in non mammalian species [123]. This evolutionary but also functional evidence suggests DUF1220 and its expansion in the human lineage is critical to higher cognitive functions.

In human, evidence for adaptive evolution has been shown by the pattern of variation in copy number of the AMY1 gene which is coding for the enzyme responsible for starch hydrolysis. This pattern has been shown to correlate positively with amylase protein levels and the population differences in starch intake, demonstrating the importance of starchy foods in human evolution [124]. Also in flies, SVs were found to affect genes and most notably high-frequency duplication CNVs were found to involve toxin-response genes (for example, Cyp6g1 contributing to resistance to DTT) suggesting positive selection on these CNVs [125].

Identifying SVs

Until recently identification and genotyping of structural variants was

performed using difficult and costly techniques such as array based comparative genome hybridization (aCGH) [126-129] or fosmid paired end sequencing (FPES) [130]. The former technique involves DNA microarrays which test the relative frequencies of probe DNA segments between two genomes and the latter is a laborious method based on low coverage Sanger style sequencing of fosmid ends.

High density SNP-arrays can provide some insight into structural variation because in many cases structural variants reveal themselves through specific signatures in SNP genotype data. In particular polymorphic deletions, which are a loss of genetic material and therefore results in aberrant patterns of SNP genotypes. Genotype mining will reveal putative deletions as large stretches of null genotypes or homozygous genotype calls deviating significantly from their expected values under Hardy-Weinberg equilibrium. Mendelian inheritance inconsistencies can arise if one parent harbors a deletion. However, such SNPs are routinely discarded as technical failures of genotyping. Other CNV calling approaches utilize probe intensity measurements from the SNP arrays rather than genotype calls [131].

Recent developments in NGS technology have dramatically economized paired end, whole genome sequencing. Mate pair or paired end sequencing methods have proven to be extremely useful for structural variant discovery [130,132,133]. In this approach, two paired reads are generated at an approximately known distance in the subject's genome. The reads are mapped to a reference genome, and pairs mapping discordant with the expected length, or with anomalous orientation, suggest structural variants. Paired end mapping (PEM) algorithms, which are based on the mining of such mate pairs, have been successfully used to discover structural variants. The combination of next generation sequencing and paired end mapping is better at detecting SVs than aCGH because it is able to detect smaller rearrangements which actually depends on the insert size used for constructing the genome library. By sequencing multiple libraries of varying insert sizes the whole size range of structural variants can be discovered.

These improvements in detection resolution has created an apparent shift downward in SV size, resulting in an explosion in reports of variants in the range of 100 bp to 1 kb [74,134,135]. These variants are smaller than the operational definition of CNV and therefore are referred to as SVs. The SVs can be further classified as insertion, deletion, inversion or translocation. Also copy invariant structural variants, such as inversions, and the exact location of variation breakpoints can be determined by sequencing [136]. Moreover,

breakpoint resolution, copy-number accuracy, specificity and sensitivity can be improved simply by increasing sequence coverage. However, detecting variation in repeated regions remains problematic because of the reliability and uniqueness of the mapping of either end of a mate pair.
The majority of large structural variants is expected to be present within these repeated regions [101,102,137] and array based comparative genome hybridization (aCGH) is still the platform of choice for the detection of variation in these regions. Also, genotyping individuals for previously known variants is currently more cost effective on array based technologies compared to NGS.
The Human Genome Structural Variation Group [103,138] and the 1000 Genomes Project [73,74] aim for characterization at the sequence level of the genomes of many individuals affected by common disorders and of subjects belonging to different ethnic groups in the next few years. In contrast to human where almost 50,000 SVs have been reported and stored in the Database of Genomic Variants [128] and of which some are linked to disease [139] only a few SVs have been detected in farm animals such as pig [140], cow [141,142], duck, turkey, and chicken [43]. In these species also some phenotypes have been linked to structural variation like the KIT locus determining the hair color in pig [143], the pea-comb phenotype in chicken [144] and late feathering in chicken [145]. Although cost-effective genotyping techniques for known SVs in animals [146,147] have been described, the large scale identification of SVs in animal genomes have not yet been published.

Aim and outline of this thesis

The aim of this thesis is to contribute to the development of a genetic variability repository for farm animals, allowing the construction of linkage maps, SNP genotyping based estimation of kinship and pedigree reconstruction and QTL studies. The developed SNP repositories also contribute to the implementation of GWA studies and GWMAS in a broad range of farm animals including turkey, ducks and pigs. We also set the first step in developing a repository for SVs, a relatively new genetic marker in animal sciences, in the chicken genome. The specific objective of the work described in this thesis was to design and validate experimental approaches and bioinformatics data pipelines to address the challenge of the cost-effective identification of genetic markers in DNA sequencing data.
In Chapter 2 we provide a strategy of how SNPs can be mined in publicly available whole genome sequencing datasets consisting of output from

traditional capillary sequencing platforms. Publicly available mapping data was included in the analysis to provide a rough estimate of the genomic location for each identified SNP.
In chapter 3 we report on the use of next generation sequencing technology in species that lack a sequenced reference genome as well as sufficient sequence repository. Samples of multiple individuals were pooled and genome complexity was systematically reduced to cost effectively detect genetic variation.
In chapter 4 we report on the application of paired end NGS to obtain a first impression about the presence of structural variation in the chicken genome. SVs were identified as abnormally aligned read pairs that have an orientation discordant from what was expected based on the constructed genome library and the chicken reference genome. We designed SV detection parameters to reliably distinguish true structural variation from false positive predictions.
In chapter 5 we anticipated on the fast developments in NGS technology and re-implemented our SNP detection approach described in chapter 3. Identified SNPs were compared with available repositories and the subset of common SNPs was validated by genotyping and used to benchmark other subsets of our SNP data.

References

1.Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT, Galver LM, Fan J, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, Bakker PIWD, Barrett J,

Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-861.

2.Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K: **Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes.** *Nat Genet* 2006, **38**:320-323.

3.Helgadottir A, Manolescu A, Helgason A, Thorleifsson G, Thorsteinsdottir U, Gudbjartsson DF, Gretarsdottir S, Magnusson KP, Gudmundsson G, Hicks A, Jonsson T, Grant SFA, Sainz J, O'Brien SJ, Sveinbjornsdottir S, Valdimarsson EM, Matthiasson SE, Levey AI, Abramson JL, Reilly MP, Vaccarino V, Wolfe ML, Gudnason V, Quyyumi AA, Topol EJ, Rader DJ, Thorgeirsson G, Gulcher JR, Hakonarson H, Kong A, Stefansson K: **A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction.** *Nat Genet* 2006, **38**:68-74.

4.Consortium BH, Gibbs RA, Taylor JF, Tassell CPV, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S, Matukumalli LK, McEwan JC, Nazareth LV, Schnabel RD, Weinstock GM, Wheeler DA, Ajmone-Marsan P, Boettcher PJ, Caetano AR, Garcia JF, Hanotte O, Mariani P, Skow LC, Sonstegard TS, Williams JL, Diallo B, Hailemariam L, Martinez ML, Morris CA, Silva LOC, Spelman RJ, Mulatu W, Zhao K, Abbey CA, Agaba M, Araujo FR, Bunch RJ, Burton J, Gorni C, Olivier H, Harrison BE, Luff B, Machado MA, Mwakaya J, Plastow G, Sim W, Smith T, Thomas MB, Valentini A, Williams P, Womack J, Woolliams JA, Liu Y, Qin X, Worley KC, Gao C, Jiang H, Moore SS, Ren Y, Song X, Bustamante CD, Hernandez RD, Muzny DM, Patil S, Lucas AS, Fu Q, Kent MP, Vega R, Matukumalli A, McWilliam S, Sclep G, Bryc K, Choi J, Gao H, Grefenstette JJ, Murdoch B, Stella A, Villa-Angulo R, Wright M, Aerts J, Jann O, Negrini R, Goddard ME, Hayes BJ, Bradley DG, Silva MBD, Lau LPL, Liu GE, Lynn DJ, Panzitta F, Dodds KG: **Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds.** *Science* 2009, **324**:528-532.

5.Botstein D, White RL, Skolnick M, Davis RW: **Construction of a genetic linkage map in man using restriction fragment length polymorphisms.** *Am J Hum Genet* 1980, **32**:314-331.

6.Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD: **Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms.** *Nature* 1988, **335**:721-726.

7.Klungland H, Våge DI, Gomez-Raya L, Adalsteinsson S, Lien S: **The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination.** *Mamm Genome* 1995, **6**:636-639.

8.Marklund L, Moller MJ, Sandberg K, Andersson L: **A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses.** *Mamm Genome* 1996, **7**:895-899.

9.Kijas JM, Moller M, Plastow G, Andersson L: **A frameshift mutation in MC1R and a high frequency of somatic reversions cause black spotting in pigs.** Genetics 2001, **158**:779-785.

10.Kerje S, Lind J, Schütz K, Jensen P, Andersson L: **Melanocortin 1-receptor (MC1R) mutations are associated with plumage colour in chicken.** *Anim Genet* 2003, **34**:241-248.

11.Milan D, Jeon JT, Looft C, Amarger V, Robic A, Thelander M, Rogel-

Gaillard C, Paul S, Iannuccelli N, Rask L, Ronne H, Lundström K, Reinsch N, Gellin J, Kalm E, Roy PL, Chardon P, Andersson L: **A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle.** *Science* 2000, **288**:1248-1251.

12. Tanksley SD: **Mapping polygenes.** *Annu Rev Genet* 1993, **27**:205-233.
13. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H, Crooijmans RPMA, Besnier F, Lathrop M, Muir WM, Wong GK, Gut I, Andersson L: **A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate.** *Genome Res* 2009, **19**:510-519.
14. Arias JA, Keehan M, Fisher P, Coppieters W, Spelman R: **A high density linkage map of the bovine genome.** BMC Genet 2009, **10**:18.
15. Vingborg RKK, Gregersen VR, Zhan B, Panitz F, Høj A, Sørensen KK, Madsen LB, Larsen K, Hornshøj H, Wang X, Bendixen C: **A robust linkage map of the porcine autosomes based on gene-associated SNPs.** BMC Genomics 2009, **10**:134.
16. Raadsma HW, Thomson PC, Zenger KR, Cavanagh C, Lam MK, Jonas E, Jones M, Attard G, Palmer D, Nicholas FW: **Mapping quantitative trait loci (QTL) in sheep. I. A new male framework linkage map and QTL for growth rate and body weight.** *Genet Sel Evol* 2009, **41**:34.
17. Hu Z, Fritz ER, Reecy JM: **AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond.** *Nucleic Acids Res* 2007, **35**:D604-9.
18. Darvasi A, Soller M: **A simple method to calculate resolving power and confidence interval of QTL map location.** *Behav Genet* 1997, **27**:125-132.
19. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
20. Schaeffer LR: **Strategy for applying genome-wide selection in dairy cattle**. *J Anim Breed Genet* 2006, **123**:218-223.
21. Bernardo R, Yu J: **Prospects for Genomewide Selection for Quantitative Traits in Maize**. *Crop Sci* 2007, **47**:1082-1090.
22. Hakonarson H, Qu H, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC, Eckert AW, Annaiah K, Lawson ML, Otieno FG, Santa E, Shaner JL, Smith RM, Onyiah CC, Skraban R, Chiavacci RM, Robinson LJ, Stanley CA, Kirsch SE, Devoto M, Monos DS, Grant SFA, Polychronakos C: **A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study.** *Diabetes* 2008, **57**:1143-1146.

23. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JRB, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch A, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin M, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJF, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CNA, Doney ASF, Morris AD, Smith GD, Hattersley AT, McCarthy MI: **A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.** *Science* 2007, **316**:889-894.
24. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JHM, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AAC, Ovington NR, Allen J, Adlem E, Leung H, Wallace C, Howson JMM, Guja C, Ionescu-Tîrgovi?te C, Finland GOT1DI, Simmonds MJ, Heward JM, Gough SCL, Consortium WTCC, Dunger DB, Wicker LS, Clayton DG: **Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.** *Nat Genet* 2007, **39**:857-864.
25. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008, 9:356-369.
26. Snelling WM, Allan MF, Keele JW, Kuehn LA, McDaneld T, Smith TPL, Sonstegard TS, Thallman RM, Bennett GL: **Genome-wide association study of growth in crossbred beef cattle.** *J Anim Sci* 2010, **88**:837-848.
27. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES: **Linkage disequilibrium in the human genome.** *Nature* 2001, **411**:199-204.
28. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225-2229.
29. Abasht B, Sandford E, Arango J, Settar P, Fulton JE, O'Sullivan NP, Hassen A, Habier D, Fernando RL, Dekkers JCM, Lamont SJ: **Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations.** *BMC Genomics* 2009, **10** Suppl 2:S2.
30. Amaral AJ, Megens H, Crooijmans RPMA, Heuven HCM, Groenen MAM:

Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 2008, **179**:569-579.
31. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome.** *Science* 2004, **304**:581-584.
32. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**:321-324.
33. Collins FS, Guyer MS, Charkravarti A: **Variations on a theme: cataloging human DNA sequence variation.** *Science* 1997, **278**:1580-1581.
34. Kruglyak L, Nickerson DA: **Variation is the spice of life.** *Nat Genet* 2001, **27**:234-236.
35. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ: **High-resolution haplotype block structure in the cattle genome.** *BMC Genet* 2009, **10**:19.
36. Consortium IH: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299-1320.
37. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations.** *Science* 2005, **307**:1072-1079.
38. Andersson L, Georges M: **Domestic-animal genomics: deciphering the genetics of complex traits.** *Nat Rev Genet* 2004, **5**:202-212.
39. Consortium ICGS: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**:695-716.
40. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Smith J, Searle SMJ: **Ensembl's 10th year.** *Nucleic Acids Res* 2010, **38**:D557-D562.
41. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Graves JAM: **The promise of comparative genomics in mammals.** *Science* 1999,

286:458-62, 479-81.
42. Wienberg J, Stanyon R: **Comparative painting of mammalian chromosomes.** *Curr Opin Genet Dev* 1997, **7**:784-791.
43. Skinner BM, Robertson LBW, Tempest HG, Langley EJ, Ioannou D, Fowler KE, Crooijmans RPMA, Hall AD, Griffin DK, Völker M: **Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis.** *BMC Genomics* 2009, **10**:357.
44. Kehrer-Sawatzki H, Cooper DN: **Structural divergence between the human and chimpanzee genomes.** Hum Genet 2007, **120**:759-778.
45. Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM: **Gene copy number variation spanning 60 million years of human and primate evolution.** *Genome Res* 2007, **17**:1266-1277.
46. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RPMA, Groenen MAM, Deryusheva S, Gaginskaya E, Carré W, Waddington D, Talbot R, Völker M, Masabanda JS, Burt DW: **Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution.** *BMC Genomics* 2008, **9**:168.
47. Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE: **Analysis of recent segmental duplications in the bovine genome.** *BMC Genomics* 2009, **10**:571.
48. Stanyon R, Rocchi M, Capozzi O, Roberto R, Misceo D, Ventura M, Cardone MF, Bigoni F, Archidiacono N: **Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres.** *Chromosome Res* 2008, **16**:17-39.
49. Li W, Freudenberg J: **Two-parameter characterization of chromosome-scale recombination rate.** *Genome Res* 2009, **19**:2300-2307.
50. Jenkins S, Gibson N: **High-throughput SNP genotyping.** *Comparative and Functional Genomics* 2002, **3**:57-66.
51. Syvänen A: **Toward genome-wide SNP genotyping.** *Nat Genet* 2005, **37** Suppl:S5-10.
52. Zhao Z, Fu Y, Hewett-Emmett D, Boerwinkle E: **Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution.** *Gene* 2003, **312**:207-213.
53. Jungerius BJ, Rattink AP, Crooijmans RPMA, Poel JJVD, Oost BAV, Pas MFWT, Groenen MAM: **Development of a single nucleotide polymorphism map of porcine chromosome 2**. *Anim Genet* 2003, **34**:429-437.
54. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: **Overlapping genomic**

sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res* 1998, **8**:748-754.
55. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
56. Gu Z, Hillier L, Kwok PY: **Single nucleotide polymorphism hunting in cyberspace.** *Hum Mutat* 1998, **12**:221-225.
57. Green P: **http://bozeman.mbt.washington.edu/phrap.docs/phrap.html.** 1996, :.
58. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
59. Nickerson DA, Tobe VO, Taylor SL: **PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing.** *Nucleic Acids Res* 1997, **25**:2745-2751.
60. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet* 1999, **23**:452-456.
61. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA: **Automating sequence-based detection and genotyping of SNPs from diploid samples.** *Nat Genet* 2006, **38**:375-381.
62. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, Rowe W, Liu PP, Gibbs RA, Buetow KH: **SNPdetector: a software tool for sensitive and accurate SNP detection.** *PLoS Comput Biol* 2005, **1**:e53.
63. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, Jonghe PD, Broeckhoven CV, Rijk PD: **novoSNP, a novel computational tool for sequence variation discovery.** *Genome Res* 2005, **15**:436-442.
64. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER: **PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data.** *Genome Res* 2007, **17**:659-666.
65. Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, Cirera S, Thomsen B, Madsen LB, Høj A, Vingborg RK, Zahn B, Wang X, Wang X, Wernersson R, Jørgensen CB, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Brunak S, Fredholm M, Bendixen C: **SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation.** *Bioinformatics* 2007, **23**:i387-i391.
66. Barker G, Batley J, Sullivan HO, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**:421-422.
67. Tang J, Vosman B, Voorrips RE, Linden CGVD, Leunissen JAM:

QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* 2006, **7**:438.

68. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases.** *Genome Res* 2001, **11**:1725-1729.
69. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Etten WJV, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, Group ISMW: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409**:928-933.
70. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656-664.
71. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7**:203-214.
72. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-876.
73. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456**:60-65.
74. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS,

Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Cooley RN, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fajardo KVF, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang G, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-59.

75. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome.** *Nat Biotechnol* 2009, **27**:847-852.

76. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858.

77. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome.** *BMC Genomics* 2008, **9**:312.

78. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T, Yiu S, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2009, :.
79. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biol* 2009, **10**:R32.
80. Wiedmann RT, Smith TPL, Nonneman DJ: **SNP discovery in swine by reduced representation and high throughput pyrosequencing.** *BMC Gene*t 2008, **9**:81.
81. Sánchez CC, Smith TPL, Wiedmann RT, Vallejo RL, Salem M, Yao J, Rexroad CE: **Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library.** *BMC Genomics* 2009, **10**:559.
82. Vos P, Hogers R, Bleeker M, Reijans M, Lee TVD, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M: **AFLP: a new technique for DNA fingerprinting.** *Nucleic Acids Res* 1995, **23**:4407-4414.
83. Orsouw NJV, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, Poel HVD, Oeveren JV, Verstegen H, Eijk MJTV: **Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes.** *PLoS One* 2007, **2**:e1172.
84. Altshuler D, Pollara VJ, Cowles CR, Etten WJV, Baldwin J, Linton L, Lander ES: **An SNP map of the human genome generated by reduced representation shotgun sequencing.** *Nature* 2000, **407**:513-516.
85. Peterson DG, Wessler SR, Paterson AH: **Efficient capture of unique**

sequences from eukaryotic genomes. *Trends Genet* 2002, **18**:547-550.
86. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**:e105.
87. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER: **Whole-genome sequencing and variant discovery in C. elegans.** *Nat Methods* 2008, **5**:183-188.
88. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ: **Identification of genetic variants using bar-coded multiplexed sequencing.** *Nat Methods* 2008, **5**:887-893.
89. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24**:713-714.
90. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
91. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
92. Li R, Yu C, Li Y, Lam T, Yiu S, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25**:1966-1967.
93. Ohno S: *Evolution by gene duplication***.** Springer Verlag, New York; 1970.
94. Hurles M: **Gene duplication: the genomic trade in spare parts.** *PLoS Biol* 2004, **2**:E206.
95. Lupski JR: **Genomic rearrangements and sporadic disease.** *Nat Genet* 2007, **39**:S43-S47.
96. Ommen GBV: **Frequency of new copy number variation in humans.** *Nat Genet* 2005, **37**:333-334.
97. Potocki L, Chen KS, Koeuth T, Killian J, Iannaccone ST, Shapira SK, Kashork CD, Spikes AS, Shaffer LG, Lupski JR: **DNA rearrangements on both homologues of chromosome 17 in a mildly delayed individual with a family history of autosomal dominant carpal tunnel syndrome.** *Am J Hum Genet* 1999, **64**:471-478.
98. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: **Molecular basis of base substitution hotspots in Escherichia coli.** *Nature* 1978, **274**:775-780.
99. Cooper DN, Youssoufian H: **The CpG dinucleotide and human genetic**

disease. *Hum Genet* 1988, **78**:151-155.
100.Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution.** *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.
101.Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80**:91-104.
102.Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-454.
103.Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.
104.McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, Bakker PIWD, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D: **Integrated detection and population-genetic analysis of SNPs and copy number variation.** *Nat Genet* 2008, **40**:1166-1174.
105.Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA: S**ystematic assessment of copy number variant detection via genome-wide SNP genotyping.** *Nat Genet* 2008, **40**:1199-1203.
106.Lee S, Cheran E, Brudno M: **A robust framework for detecting structural variations in a genome.** *Bioinformatics* 2008, **24**:i59-i67.
107.Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J,

Hulihan M, Peuralinna T, Dutra A, Nussbaum R, Lincoln S, Crawley A, Hanson M, Maraganore D, Adler C, Cookson MR, Muenter M, Baptista M, Miller D, Blancato J, Hardy J, Gwinn-Hardy K: **alpha-Synuclein locus triplication causes Parkinson's disease.** *Science* 2003, **302**:841.

108.Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King M, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M: **Strong association of de novo copy number mutations with autism.** *Science* 2007, **316**:445-449.

109.Hollox EJ, Huffmeier U, Zeeuwen PLJM, Palla R, Lascorz J, Rodijk-Olthuis D, Kerkhof PCMVD, Traupe H, Jongh GD, Heijer MD, Reis A, Armour JAL, Schalkwijk J: **Psoriasis is associated with increased beta-defensin genomic copy number.** *Nat Genet* 2008, **40**:23-25.

110.McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, Jones PBB, McLean L, O'Donnell JL, Pokorny V, Spellerberg M, Stamp LK, Willis J, Steer S, Merriman TR: **Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis.** *Ann Rheum Dis* 2008, **67**:409-413.

111.Lupski JR, Stankiewicz P: **Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.** *PLoS Genet* 2005, **1**:e49.

112.Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, Ledbetter DH, Greenberg F, Patel PI: **Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A.** *Nat Genet* 1992, **1**:29-33.

113.Chance PF, Alderson MK, Leppig KA, Lensch MW, Matsunami N, Smith B, Swanson PD, Odelberg SJ, Disteche CM, Bird TD: **DNA deletion associated with hereditary neuropathy with liability to pressure palsies.** *Cell* 1993, **72**:143-151.

114.Nathans J, Piantanida TP, Eddy RL, Shows TB, Hogness DS: **Molecular genetics of inherited variation in human color vision.** *Science* 1986, **232**:203-210.

115.Lifton RP, Dluhy RG, Powers M, Rich GM, Cook S, Ulick S, Lalouel JM: **A chimaeric 11 beta-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension.** *Nature* 1992, **355**:262-265.

116.Velagaleti GVN, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, Withers M, Lupski JR, Stankiewicz P: **Position effects due to**

chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *Am J Hum Genet* 2005, **76**:652-662.

117.Kleinjan DA, Heyningen VV: **Long-range control of gene expression: emerging mechanisms and disruption in disease.** *Am J Hum Genet* 2005, **76**:8-32.

118.Kurotaki N, Shen JJ, Touyama M, Kondoh T, Visser R, Ozaki T, Nishimoto J, Shiihara T, Uetake K, Makita Y, Harada N, Raskin S, Brown CW, Höglund P, Okamoto N, Lupski JR: **Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency.** *Genet Med* 2005, **7**:479-483.

119.Yan J, Bi W, Lupski JR: **Penetrance of craniofacial anomalies in mouse models of Smith-Magenis syndrome is modified by genomic sequence surrounding Rai1: not all null alleles are alike.** *Am J Hum Genet* 2007, **80**:518-525.

120.Duncan IW: **Transvection effects in Drosophila.** *Annu Rev Genet* 2002, **36**:521-556.

121.Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK: **A high-resolution survey of deletion polymorphism in the human genome.** *Nat Genet* 2006, **38**:75-81.

122.Dopman EB, Hartl DL: **A portrait of copy-number polymorphism in Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 2007, **104**:19920-19925.

123.Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM: **Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains.** *Science* 2006, **313**:1304-1307.

124.Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC: **Diet and the evolution of human amylase gene copy number variation.** *Nat Genet* 2007, **39**:1256-1260.

125.Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M: **Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster.** *Science* 2008, **320**:1629-1631.

126.Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: **High resolution analysis of DNA copy number variation using**

comparative genomic hybridization to microarrays. *Nat Genet* 1998, **20**:207-211.

127.Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41-46.

128.Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36**:949-951.

129.Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: **Large-scale copy number polymorphism in the human genome.** *Science* 2004, **305**:525-528.

130.Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37**:727-732.

131.Yau C, Holmes CC: **CNV discovery using SNP genotyping arrays.** *Cytogenet Genome Res* 2008, **123**:307-312.

132.McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, Vega FMDL, Blanchard AP: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19**:1527-1541.

133.Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318**:420-426.

134.Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A,

Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PAW, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40**:722-729.

135. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20**:265-272.
136. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing.** *Nat Methods* 2009, **6**:S13-S20.
137. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77**:78-88.
138. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, Lupski JR, Mullikin JC, Pritchard JK, Sebat J, Sherry ST, Smith D, Valle D, Waterston RH: **Completing the map of human genetic variation.** *Nature* 2007, **447**:161-165.
139. Wain LV, Armour JAL, Tobin MD: **Genomic copy number variation, human health, and disease.** *Lancet* 2009, **374**:340-350.
140. Fadista J, Nygaard M, Holm L, Thomsen B, Bendixen C: **A snapshot of CNVs in the pig genome.** *PLoS One* 2008, **3**:e3916.
141. Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE: **Analysis of recent segmental duplications in the bovine genome.** *BMC Genomics* 2009, **10**:571.
142. Liu GE, Tassel CPV, Sonstegard TS, Li RW, Alexander LJ, Keele JW, Matukumalli LK, Smith TP, Gasbarre LC: **Detection of germline and somatic copy number variations in cattle.** *Dev Biol (Basel)* 2008, **132**:231-237.
143. Moller MJ, Chaudhary R, Hellmén E, Höyheim B, Chowdhary B, Andersson L: **Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor.** *Mamm Genome* 1996, **7**:822-830.
144. Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin C, Imsland F, Hallböök F, Andersson L: **Copy**

number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. *PLoS Genet* 2009, **5**:e1000512.

145.Elferink MG, Vallée AAA, Jungerius AP, Crooijmans RPMA, Groenen MAM: **Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken.** *BMC Genomics* 2008, **9**:391.

146.Seo B, Park E, Ahn S, Lee S, Kim J, Im H, Lee J, Cho I, Kong I, Jeon J: **An accurate method for quantifying and analyzing copy number variation in porcine KIT by an oligonucleotide ligation assay.** *BMC Genet* 2007, **8**:81.

147.Pielberg G, Day AE, Plastow GS, Andersson L: **A sensitive method for detecting variation in copy numbers of duplicated genes.** *Genome Res* 2003, **13**:2171-2177.

2 Mining for single nucleotide polymorphisms in pig genome sequence data

Hindrik HD Kerstens[1], Sonja Kollers[2], Arun Kommadath[1], Marisol del Rosario[1], Bert Dibbits[1] Sylvia M Kinders[1], Richard P Crooijmans[1] Martien AM Groenen[1§]

1Animal Breeding and Genetics Group, Wageningen University, P.O. Box 9101, Wageningen, 6701 BH, The Netherlands
2IPG, Institute for Pig Genetics, P.O. Box 43,Beuningen, 6640 AA, The Netherlands
§Corresponding author

Email addresses:
HHDK: hindrik.kerstens@wur.nl
SK: sonjakollers@gmx.de
AK: arun.kommadath@wur.nl
MdR: marisol_del_rosario@hotmail.com
BD: bert.dibbits@wur.nl
Smk: sylvia.kinders@wur.nl
RPC: richard.crooijmans@wur.nl
MAMG: martien.groenen@wur.nl

Published in BMC Genomics. 2009 Jan 6;10:4.

Abstract

Background

Single nucleotide polymorphisms (SNPs) are ideal genetic markers due to their high abundance and the highly automated way in which SNPs are detected and SNP assays are performed. The number of SNPs identified in the pig thus far is still limited.

Results

A total of 4.8 million whole genome shotgun sequences obtained from the NCBI trace-repository with center name "SDJVP", and project name "Sino-Danish Pig Genome Project" were analysed for the presence of SNPs. Available BAC and BAC-end sequences and their naming and mapping information, all obtained from SangerInstitute FTP site, served as a rough assembly of a reference genome. In 1.2 Gb of pig genome sequence, we identified 98,151 SNPs in which one of the sequences in the alignment represented the polymorphism and 6,374 SNPs in which two sequences represent an identical polymorphism. To benchmark the SNP identification method, 163 SNPs, in which the polymorphism was represented twice in the sequence alignment, were selected and tested on a panel of three purebred boar lines and wild boar. Of these 163 in silico identified SNPs, 134 were shown to be polymorphic in our animal panel.

Conclusions

This SNP identification method, which mines for SNPs in publicly available porcine shotgun sequences repositories, provides thousands of high quality SNPs. Benchmarking in an animal panel showed that more than 80% of the predicted SNPs represented true genetic variation.

Background

Single nucleotide polymorphisms (SNPs), one of the most abundant types of sequence polymorphisms in the genome, are the most suitable markers for genetic linkage mapping, fine-mapping and haplotype reconstruction. Over the past decade, SNPs have been the marker of choice due to their high stability, density and the highly automated way in which SNPs are detected and SNP assays are performed. However only a limited number of SNPs have been identified in the pig, a species of considerable economical and medical importance. A few thousand SNPs in the pig are currently available, and these were mainly identified in expressed genes by either in vitro techniques [1] or by

mining porcine expressed sequence tag (EST) sequence databases [2,3]. In humans, the large-scale identification and characterization of SNPs has attracted much more attention, and consequently over 14 million SNPs (dbSNP build 128) have been identified [4], 3.1 million of which have been genotyped SNPs [5] and the SNP density is estimated as one SNP per 1000-2000 bases [6]. Genome scans with high SNP densities have proven to be an effective tool in whole genome association studies to identify genes involved in complex genetic traits [7,8,9,10]. The SNP density in pigs is about four-fold higher than that in

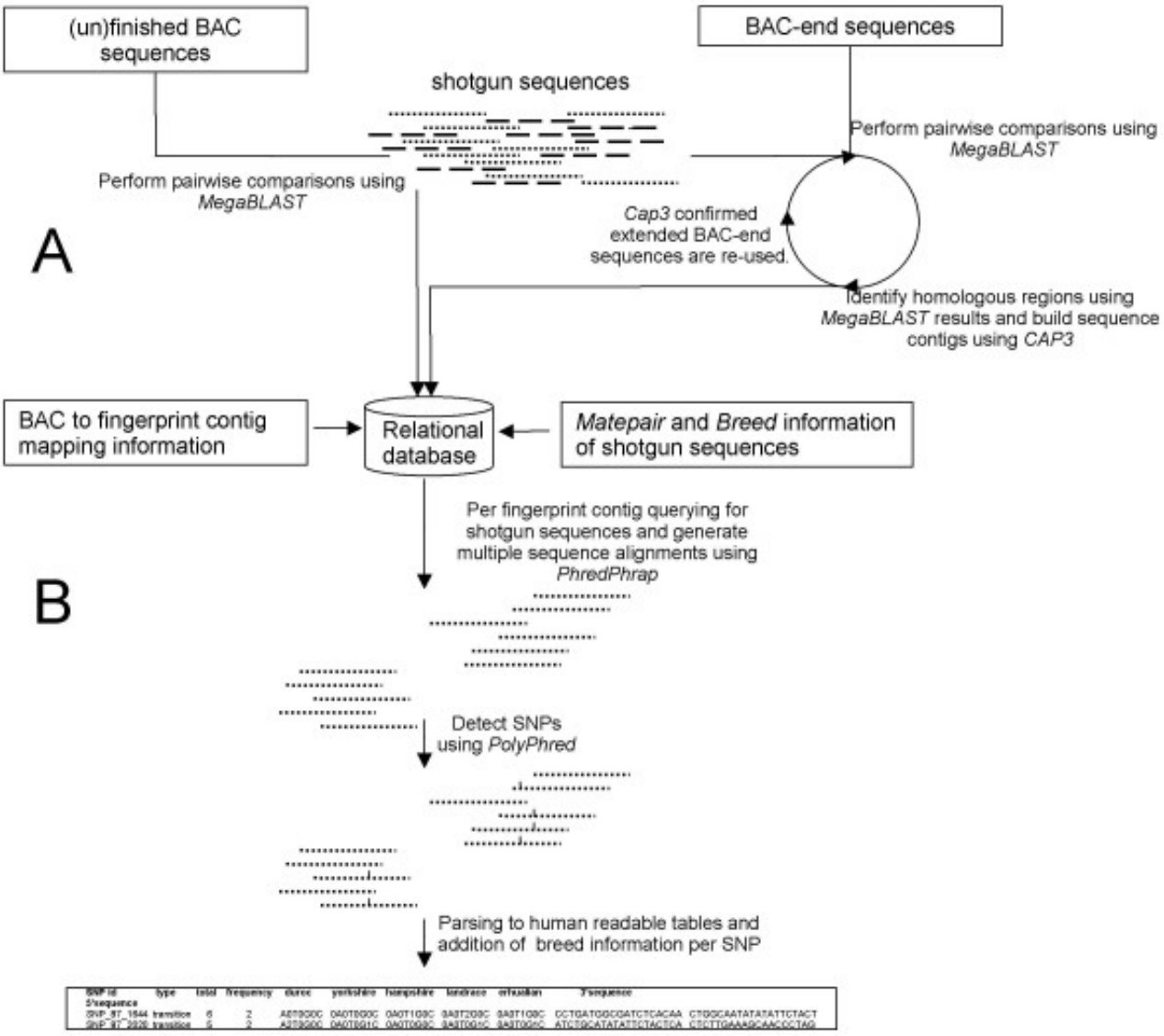

Figure 1: Steps performed for SNP mining in whole genome shotgun sequences in which BAC(-end) sequences and their mapping information served as a reference genome. Initially shotgun sequences are assigned to fingerprint contigs (A). Subsequently per fingerprint contig SNPs were mined using PhredPhrap and PolyPhred (B).

humans with SNPs found at, on average, every 300 to 400 bps [11]. Despite the availability of the most highly continuous bacterial artificial chromosome (BAC) map of any mammalian genome [12] and the ongoing sequencing efforts in the pig [13], no large scale SNP mining on pig genome sequences has been published. The lack of a pig genome draft assembly still hampers the traditional method of identifying SNPs, in which DNA shotgun sequences of different individuals are aligned to a genomic region of interest using alignment algorithms [14]. In these alignments, sequences are easily compared and SNP candidates can be reliably detected by computational methods like PolyPhred [15], which has been extensively tested for human SNP discovery [16,17,18]. Despite the unavailability of a draft sequence of the pig genome, a wealth of high quality sequence and mapping data is publicly available that can be used for SNP detection purposes.

Here we describe a high throughput genome sequence mining pipeline from data of the ongoing pig genome sequencing project. With this approach, we performed a SNP mining analysis on the whole genome shotgun dataset generated by the Danish-Chinese Pig Genome Sequencing Initiative [19] that is publicly available in the NCBI Trace Archive. BAC sequence data and the BAC mapping information to the porcine physical map [12] were combined and we used this as a crude assembly of a reference genome sequence. The pipeline is built from existing public software packages and implemented on a computer cluster, which enables efficient mining of large sequence data sets in parallel.

The encouraging outcome of this study is a good starting point for the development of a rapidly growing genome-wide set of SNP markers in the pig.

Results

Clustering

At completion of this analysis, the number of finished and contigs of unfinished porcine BAC sequences was 318 and 84,017, resulting in 50,225,986 and 1,164,409,065 total nucleotides, respectively. The NCBI Trace repository contained 4,774,371 whole genome shotgun sequences for center SDJVP, with a total of 3,478,199,073 nucleotides.

Because the analysis of the complete data set for the whole genome was computationally too demanding, the identification of SNPs was performed by a 2-step process. First, the shotgun sequences were assigned to a fingerprint contig by clustering based on their sequence similarity to BAC and BAC-end sequences. The results of the clustering by alignment were stored in a relational

database. The BAC and the BAC-end naming as well as the mapping data provided the necessary information to assign the obtained sequence clusters to a specific fingerprint contig on the porcine physical BAC map. Clustering of the shotgun reads with BAC or BAC-end sequences is outlined in Figure 1A.
This approach enabled the chromosomal assignment of the sequences, even for chromosomes and chromosomal regions for which currently no assembled chromosome sequence is available at the pre-ensemble [20] website. In total, 838,711 shotgun sequences were clustered and assigned to a specific fingerprint contig (fpc) and 97.7% of these shotgun sequences mapped to a single unique fpc.

Identifying candidate SNPs

In the second step, the actual identification of SNPs was performed per fingerprint contig. In this respect, a fingerprint contig can be considered a 'genomic region of interest', which is the starting point in traditional SNP mining in species for which a genome draft is available. Per fingerprint contig, the relational database was queried for shotgun reads, in which repetitive sequences were tagged, and were aligned using PhredPhrap [21,22,23]. Finally, the alignments were searched for SNPs using PolyPhred [15] as outlined in Figure 1B. Identified SNPs were categorised by the number (one, two, three or four) of sequences that represent identical nucleotide substitution on the SNP position in the sequence alignment. SNP prediction results of all fingerprint contigs were combined and analysed for redundancy. Redundancy was expected, because a small fraction (2.3%) of the shotgun reads did not uniquely map to a single fingerprint contig. Paralogous and repetitive sequences typically cause ambiguous clustering results. Although the initial clustering of shotgun sequences was refined in the alignment procedure by Phrap [23], a small number of SNPs still mapped to two distinct genomic regions. These

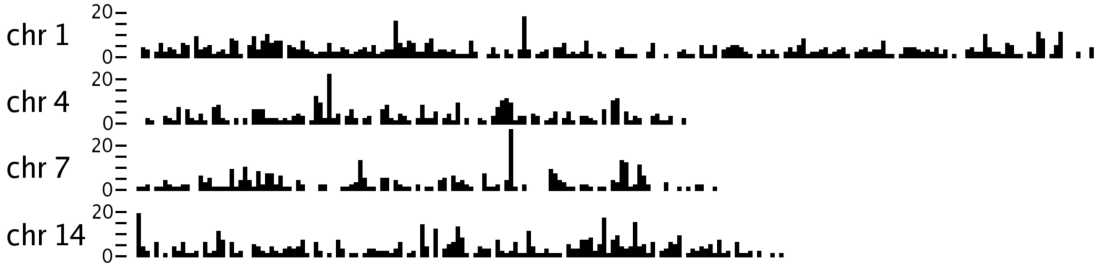

Figure 2: Distribution of SNPs on pig chromosomes 1, 4, 7 and 14. The X-axis represents the chromosome in intervals of 1Mb in size. On the Y-axis the number of identified SNPs is shown for the 1Mb intervals, each tick is five.

ambiguous SNPs were removed from the data, resulting in a final list of 98,151 unique SNPs (Table 1). The number of identified SNPs was drastically reduced when the constraint for the number of sequences representing identical nucleotide substitution in a SNP was increased. When this number was raised above two, the majority of predicted SNPs were located within a genomic context that was tagged as repetitive sequence.

Table 1: SNPs identified, substitution ratios and the fraction in repetitive context at increasing polymorphism representation constraints.

SNP representation	*Total SNPs identified*	*Transition/ transversion*	*Fraction SNPs in repetitive sequence*
1	98151	1.9	0.39
2	6374	2.8	0.60
3	1202	4.2	0.90
4	462	5.8	0.96

Distribution of SNPs over the pig genome

At completion of this analysis (Dec 2007), the sequencing of the pig genome was ongoing and most assembled BAC contig sequences were available for chromosomes 1, 4, 7, and 14 The number of SNPs as a percentage of total number of identified SNPs per analysed chromosome is provided in Table 2.

Table 2: Distribution of SNPs over the analysed pig chromosomes in percentages of the total number identified.

SNP	*Chromosome*															
representation	*1*	*2*	*3*	*4*	*5*	*7*	*8*	*9*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*
1	20,2	3,0	2,2	12,7	2,9	12,7	2,4	1,0	5,9	0,3	6,2	15,5	6,2	2,2	5,1	1,4
2	22,6	2,6	2,0	12,1	2,2	14,1	2,0	1,0	4,4	0,2	5,7	16,5	7,4	1,8	4,0	1,3

To evaluate if the SNPs distribute equally throughout the pig chromosomes the exact locations of unique SNPs predicted on chromosome 1,4,7,14 were determined by alignment. A total of 1783 SNPs that mapped uniquely were plotted along these chromosomes as shown in Figure 2.

Analysis of base changes

The SNPs in the subsets of candidate SNPs in which identical nucleotide substitution is represented in one, two, three or four sequences in the alignment were categorized according to nucleotide substitutions: C/T or G/A (transitions) and C/G, A/G, C/A, T/G (transversions). For each category, we calculated the relative nucleotide substitution frequencies for our SNP dataset and for the genomic porcine SNPs recorded in dbSNP [4] (Table 3). For the SNP subset in which identical nucleotide substitution is represented twice in the alignment, we observed a very similar relative increase in the proportion of transitions over transversions compared to the SNPs in dbSNP [4].

Table 3: Comparison of substitution frequencies of SNPs deposited in dbSNP[4] and polymorphisms identified in shotgun sequences.

			Shotgun sequence analysis			
	dbSNP (genomic)		***SNP redundancy = 1***		***SNP redundancy = 2***	
Transitions	5404	73,04%	64167	65,38%	4676	73,36%
Transversions	1995	26,96%	33984	34,62%	1698	26,64%
Total	7399		98151		6374	

SNPs in common with dbSNP

To estimate whether SNPs predicted by our method are already present in the public database of dbSNP [4], we compared the two datasets by clustering. In dbSNP [4], we selected genomic SNPs (class=1) with at least 50 bases of sequence on each side. These 7,896 SNPs were trimmed to have exactly 50 bases of flanking sequence and were analyzed for redundant records. The confirmed 7,586 unique SNPs were compared to our 98,151 predicted SNPs (singe representation of nucleotide substitution in alignment) by clustering. No clusters were formed, indicating that our dataset and dbSNP [4] share no SNPs in common.

Experimental validation of candidate SNPs

To balance the sequence context and the number of times a polymorphism is represented in the sequence alignment, SNPs in which a nucleotide substitution was represented at least twice in the sequence alignment were chosen for experimental validation. A total of 163 selected candidate SNPs were validated by genotyping in a panel of three purebred boar lines (+ wild boar). A total of 61,777 genotype analyses were performed providing, in addition to SNP prediction validation, insights into allele frequencies that will be valuable

information for association mapping and QTL studies. To measure the performance of our analyses, validated SNPs were included that previously had been used within the European Union (EU) pig biodiversity project II (Pig-BioDiv II) [24] as well as SNPs described by Rohrer et al [25] (Table 4). Also, 16 known SNPs in the IGF2-region and 14 SNPs described in a number of publications were included see[http://www.biomedcentral.com/content/supplementary/1471-2164-10-4-s1.doc]. For all 331 SNPs, the allelic variation was determined in our animal panel. In 29 cases, the predicted candidate SNPs turned out to be monomorphic. Smaller fractions of SNPs are observed to be monomorphic in the PigBiodiv and Various Literature SNP sets. The SNPs described by Rohrer et al [25] and the IGF2-region were all polymorphic in our animal panel.
For each predicted candidate SNP that appeared to be polymorphic in our panel, minor allele frequencies per boar line and overall average minor allele frequencies were calculated see [http://www.biomedcentral.com/content/supplementary/1471-2164-10-4-s2.xls].

Table 4: The performance statistics for each source of SNPs tested in our animal panel.

SNP source	*Total*	*Fraction monomorph*
PigBioDiv	99	0,07
Rohrer et al.	39	0
Various Literature	14	0,07
IGF2-region	16	0
This study	163	0,18

Discussion

Because of their highly automated high-throughput assays, SNPs are the marker of choice for molecular genetic analysis. SNPs can be obtained cost effectively by analysing public sequence data sets [26,27,28]. When sequence trace files are involved at the identification of SNPs, true polymorphisms can be distinguished from sequencing errors. Polymorphisms in which the identified base is doubtful due to a high error probability in the trace file, and therefore the most probable cause of the observed variation, are filtered out [29,30,31]. The number of sequences in which a polymorphism is represented provides information as to whether a predicted SNP represents a true polymorphism. By

filtering the observed sequence variation for polymorphisms in which the minor allele is represented at least twice in the sequence alignment, the chance that the predicted SNP is caused by sequencing errors is extremely small. Because the dataset used in our analysis consisted of shotgun sequences providing a 0.66X coverage, the sequence redundancy in our dataset is limited. This low genome coverage made it likely to detect true genetic variation already at a low sequence depth. Even SNPs with a single representation in the sequence alignment might represent true nucleotide polymorphism at this low genome coverage. However, the chance that SNPs with a single representation in the sequence alignment turns out to be monomorphic in a genotyping assay is relatively high. In order to obtain a set of high quality SNPs, we raised the threshold to a two times representation of a nucleotide substitution in the sequence alignment. A further increase of the representation constraint at this low genome coverage would lead to a SNP set in which the majority of genetic variation being detected is located in repetitive sequences. In these repetitive sequences, the degree of periodicity in nucleotide usage is high, making it hard to distinguish true allelic variation from predicted sequence variation caused by paralogous sequences. The over-representation of SNPs in repetitive sequences can be explained by errors in clustering paralogous repetitive sequences,as wel as by the 1.8 times higher SNP density in periodic DNA, which is observed in humans [32].

Although sequence quality scores and a redundancy-based approach were used to filter sequencing errors from true nucleotide polymorphisms, a non-random distribution of polymorphisms might occur in a particular dataset. These artefacts become visible when SNP statistics are compared to other SNP collections in the same species and are comparable to those found in related species. When compared to porcine SNPs deposited in dbSNP [4], our predicted SNPs in which a nucleotide substitution is represented at least twice in the sequence alignment show a similar transition/transversion ratio (Table 2). However, the transition frequency in humans was determined to be 60 to approximately 66% in vivo [16,6] and 60%-69% in silico [27,29], respectively. According to the SNP statistics in Table 1, it is evident that the transition/transversion ratio is highly biased by the fraction of SNPs in repetitive sequences in a particular dataset. A similar transition/transversion ratio for porcine SNPs deposited in dbSNP and our subset of SNPs, in which nucleotide substitutions are represented at least two times, is more likely explained by coincidence than being representative of the pig genome. The 0.6 fraction of sequences tagged as being repetitive in our SNP subset has likely

influenced the transition/transversion ratio. Therefore the transition/transversion ratio observed in the total number of predicted SNPs, single redundancy, is likely more representative for the whole pig genome. This suggests a comparable transition/transversion ratio between humans and pigs, which was expected because of the evolutionary relatedness of these species.
A comparison of our collection of predicted candidate SNPs to the porcine SNPs in dbSNP [4] revealed no SNPs in common, not to our surprise. The average SNP density in the 2.7 Gb pig genome is estimated to be one in 336 base pairs [11], indicating that only a small fraction of the expected total of tens of millions of SNPs has been identified in the pig.
Not all predicted candidate SNPs turned out to be polymorphic in the animal panel. This doesn't implicitly mean that this 0.18 fraction (Table 4) includes falsely predicted polymorphisms. SNPs in the PigBioDiv [24] and the SNPs derived from various literature [see http://www.biomedcentral.com/content/supplementary/1471-2164-10-4-s1.doc] that were previously experimentally validated resulted in (0.07) fractions of monomorphic SNPs. These fractions of monomorphic SNPs observed in this study can be explained by difference in selection of the animal panel on which the SNPs have been validated and the animal panel we used, as well as the absence of Chinese breed genetic background, near absence of Meishan and the use of another Large White in our panel.
Within our breed panel, we observed very low (<5%) Minor Allele Frequencies (MAF) in predicted candidate SNPs [see http://www.biomedcentral.com/content/supplementary/1471-2164-10-4-s2.xls].] and in the IGF2-region (data not shown). For SNPs in the IGF2-region, these low MAF are the result of intensive selection on that genomic region, whereas for the predicted candidate SNPs we did not know what to expect because of the unknown genomic location of these SNPs. Intensive selection also might have caused these very low MAF.

Conclusions

The overall performance of the SNPs identified by our genome shotgun sequence mining approach is comparable to those available in existing SNP repositories. In perspective of the ongoing sequencing of the pig genome, the SNP data generated by this approach will provide a growing number of available markers that can be applied for genotyping and will increase the SNP marker density on the pig genome.

Methods

DNA sequence data

The entire genome shotgun sequences used in this study were downloaded from the NCBI Trace repository (species SUS SCROFA, center SDJVP). For all sequences, breed and mate pair information was obtained and stored in a relational database. Finished and unfinished BAC sequences obtained within the porcine genome sequencing project were retrieved from the SangerInstitute FTP site at ftp://ftp.sanger.ac.uk/pub/sequences/pig/.
BAC-end sequences were downloaded from the Ensembl [20] FTP site at ftp://ftp.ensembl.org/pub/traces/sus_scrofa/fasta/.
BAC naming and mapping data were obtained from ftp://ftp.sanger.ac.uk/pub/S_scrofa/master_porcine_R7.tar.gz
and ftp://ftp.sanger.ac.uk/pub/S_scrofa/PIGendreads030105.txt.gz, respectively. Naming and mapping data were stored in a local relational database.

Clustering and alignment

Whole genome shotgun sequences were masked for mammalian-specific repeats and low complexity regions using RepeatMasker version open-3.1.7 [33] with options -xsmall, -species pig, default sensitivity and using the RepeatMasker Database release 20071204.
Clustering of data was performed by aligning the whole genome shotgun reads to the BAC sequences and BAC-end sequences using MegaBlast 2.2.16 [34].
Shotgun reads were aligned to BAC sequences using the alignment parameters -U T -s 122 -p95 -F m. Results were filtered for alignments with more than 90% of the shotgun sequence length. To reduce the amount of ambiguous results in the clustering, only alignment results with a bitscore >90% of the best scoring alignment for that shotgun sequence were stored in a relational database.
Clustering of shotgun sequences by alignment to BAC-end sequences was followed by assembling each cluster using CAP3 [35]. MegaBlast [34] parameters (-p 95 -s 32 -F m -U T) were matched to the CAP3 [35] settings (-o40, -p95), allowing only perfect assembled clusters. BAC-end sequences that were extended by shotgun sequences in the previous step were again used to cluster other shotgun reads until no extension occurred. Clustering results were stored in a local relational database.
Using the BAC and BAC-end naming and mapping information, we were able to query our clustering results by fingerprint contig name as used in the porcine physical map provided by the Sanger Institute. Per fingerprint contig the

shotgun sequences that clustered to this region were selected and mate pairs were added using the mate pair information in the NCBI Trace repository.
Multiple sequence alignments of the selected shotgun sequences and their mate pairs were generated by the sequence assembly script PhredPhrap [21,22,23]. Shotgun sequence trace files were used as input for PhredPhrap [21,22,23], which was run using the default parameters.

SNP identification

For identification of SNPs in the multiple sequence alignments of the shotgun sequences, we used PolyPhred [15] version 6.11 with options -snp hom -f 50, which lists homozygous SNPs with 50 bp flanking sequence.
Polyphred results were parsed into tables, information from which breed a sequence was derived and whether the SNP is located within a suspected repetitive sequence was added.

Elimination of redundancy in identified SNPs by clustering

To remove any redundancy in our SNP predictions, the results (SNP position flanked by 50 bp genomic sequence) were first stored in FASTA format. The actual clustering was performed using blastclust [36] with parameters -S 99.5 -L 1.0 -b T -p F -F F. These parameters were also used to compare our SNP prediction results to public SNPs in dbSNP [4].

Distribution of SNPs over pig chromosomes

Unique SNPs flanked by 50 bp genomic sequence predicted on chromosome 1,4,7,14 were mapped on the corresponding chromosomal sequence as provided by pre-ensemble [19]. The alignment was performed using BLAT [37] with the default parameters. SNPs that aligned uniquely to the chromosome with at least a 0.9 fraction of the flanking sequence involved in the alignment and a minimal sequence similarity of 96% were used to generate a SNP distribution plot.

SNP validation

For SNP validation, 163 SNPs were selected, with regions covered by at least 4 reads and with a minimum SNP redundancy score of 2. These SNPs were subsequently genotyped in an animal panel consisting of three purebred boar lines that originated from (1) Duroc and Belgian Landrace, (2) Large White, (3) German Pietrain and (4) Wild Boar. The four lines included 129, 120, 109 and 21 individuals, respectively. Genotyping was performed using the Illumina GoldenGate(R) Genotyping assay on an Illumina® BeadStation with veraCode(TM) technology. Oligonucleotides were designed, synthesized and assembled into oligo pooled assays (OPA) by Illumina Inc. Typing was carried out in a multiplex reaction, which included 384 loci.

Availability and requirements

The SNPs identified in this study, in which the polymorphism was represented twice in the sequence alignment, have been deposited in the National Center of Biotechnology (NCBI) SNP database (dbSNP) under submitter handle WU_ABGC. NCBI_ss 106817370–106823609 represent predicted SNPs that were not tested on in our animal panel. Predicted SNPs that were confirmed are listed in [see http://www.biomedcentral.com/content/supplementary/1471-2164-10-4-s2.xls]. SNPs with a single redundancy will be available on request.

Authors' contributions

HHDK designed and developed the SNP prediction method and wrote the manuscript. AK and MdR designed and implemented the relational database. BD, SMK, RPC and SK collected and prepared the samples and performed the genotyping analysis. SK summarized the genotyping results. MAMG coordinated and supervised the experiment implementation, and assisted in the manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

We thank Jack Leunissen, Peter Groenen and Mari Smits for critically reading the manuscript and for their helpful comments. This study was funded by Ministry of Economic Affairs IS054062 and the Institute for Pig Genetics (IPG), the Netherlands.

References

1. Fahrenkrug SC, Freking BA, Smith TPL, Rohrer GA, Keele JW: **Single nucleotide polymorphism (SNP) discovery in porcine expressed genes.** *Anim Genet* 2002, **33**:186-195.
2. Uenishi H, Eguchi-Ogawa T, Shinkai H, Okumura N, Suzuki K, Toki D, Hamasima N, Awata T: **PEDE (Pig EST Data Explorer) has been expanded into Pig Expression Data Explorer, including 10 147 porcine full-length cDNA sequences.** *Nucleic Acids Res* 2007, **35**:D650-D653.
3. Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, Cirera S, Thomsen B, Madsen LB, Høj A, Vingborg RK, Zahn B, Wang X, Wang X, Wernersson R, Jørgensen CB, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Brunak S, Fredholm M, Bendixen C: **SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation.** *Bioinformatics* 2007,13:i387-i391.
4. Sherry ST, Ward M, Sirotkin K: **dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.** *Genome Res* 1999, **9**:677-679.

5. Frazer IHCA, Ballinger DG, R.Cox D, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT, Galver LM, Fan J, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, Bakker PIWd, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-861.
6. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux

J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280**:1077-1082.

7. Trikka D, Fang Z, Renwick A, Jones SH, Chakraborty R, Kimmel M, Nelson DL: **Complex SNP-based haplotypes in three human helicases: implications for cancer association studies.** *Genome Res* 2002, **12**:627-639.
8. Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, Kirby A, Daly MJ, Jager PLD, Walsh E, Lander ES, Rioux JD, Hafler DA, Ivinson A, Rimmler J, Gregory SG, Schmidt S, Pericak-Vance MA, Akesson E, Hillert J, Datta P, Oturai A, Ryder LP, Harbo HF, Spurkland A, Myhr K, Laaksonen M, Booth D, Heard R, Stewart G, Lincoln R, Barcellos LF, Hauser SL, Oksenberg JR, Kenealy SJ, Haines JL, Consortium IMSG: **A high-density screen for linkage in multiple sclerosis.** *Am J Hum Genet* 2005, **77**:454-467.
9. Schmidt S, Pericak-Vance MA, Sawcer S, Barcellos LF, Hart J, Sims J, Prokop AM, Walt Jvd, DeLoa C, Lincoln RR, Oksenberg JR, Compston A, Hauser SL, Haines JL, Gregory SG, Group MSG: **Allelic association of sequence variants in the herpes virus entry mediator-B gene (PVRL2) with the severity of multiple sclerosis.** *Genes Immun* 2006, **7**:384-392.
10. Consortium WTCC, (TASC) ASC, Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung H, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, Clair DS, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskivina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CAB, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Genetics BiR, Committee GSS(S, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DPM, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JRB, Shields BM, Weedon MN,

Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AVS, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SCL, Seal S, (UK) BCSC, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Newport M, Sirugo G, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghori MJR, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widden C, Withers D, Cardin NJ, Davison D, Ferreira T, Pereira-Gale J, Hallgrimsdo'ttir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Brown MA, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AVS, Parkes M, Pembrey M, Stratton MR, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SHS, McGinnis R, Keniry A, Deloukas P, Reveille JD, Zhou X, Sims A, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M: **Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants.** *Nat Genet* 2007, **11**:1329-1337.

11. Jungerius BJ, Gu J, Crooijmans RPMA, Poel JJvd, Groenen MAM, Oost BAv, Pas MFWt: **Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome.** *Anim Biotechnol* 2005, **16**:41-54.
12. Humphray SJ, Scott CE, Clark R, Marron B, Bender C, Camm N, Davis J, Jenks A, Noon A, Patel M, Sehra H, Yang F, Rogatcheva MB, Milan D, Chardon P, Rohrer G, Nonneman D, Jong Pd, Meyers SN, Archibald A, Beever JE, Schook LB, Rogers J: **A high utility integrated map of the pig genome.** *Genome Biol* 2007, **8**:R139.
13. Schook LB, Beever JE, Rogers J, Humphray S, Archibald A, Chardon P, Milan D, Rohrer G, Eversole K: **Swine Genome Sequencing Consortium (SGSC): A Strategic Roadmap for Sequencing The Pig Genome.** *Comparative and Functional Genomics* 2005, **6**:251-255.
14. Myers EW, Miller W: **Optimal alignments in linear space.** *Comput Appl Biosci* 1988, **4**:11-17.
15. Nickerson DA, Tobe VO, Taylor SL: **PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing.** *Nucleic Acids Res* 1997, **25**:2745-2751.
16. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF: **DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene.** *Nat Genet* 1998, **19**:233-240.
17. Rieder MJ, Taylor SL, Tobe VO, Nickerson DA: **Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome.** *Nucleic Acids Res* 1998, **26**:967-973.

18. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA: **Automating sequence-based detection and genotyping of SNPs from diploid samples.** *Nat Genet* 2006, **38**:375-381.
19. Rasmus W, Mikkel S, Frank J, Jan G, Frank P, Hans-Henrik S, Ole C, Thomas M, Henrik H, Ami K, Jun W, Bin L, Songnian H, Wei D, Wei L, Gane W, Jun Y, Jian W, Christian B, Merete F, Soren B, Huanming Y, Lars B: **Pigs in sequence space: A 0.66X coverage pig genome survey based on shotgun sequencing.** *BMC Genomics* 2005, **6**:70.
20. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36**:D707-D714.
21. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
22. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
23. Green P: **Phrap** [http://www.phrap.orghttp://www.phrap.org]
24. Ollivier L, Alderson L, Gandini GC, Foulley JL, Haley CS, Joosten R, Rattink AP, Harlizius B, Groenen MAM, Amigues Y, Boscher MY, Russell G, Law A, Davoli R, Russo V, Matassino D, Désautés C, Fimland E, Bagga M, Delgado JV, Vega-Pla JL, Martinez AM, Ramos AM, Glodek P, Meyer JN, Plastow GS, Siggens KW, Archibald AL, Milan D, San Cristobal M, Laval G, Hammond K, Cardellino R, Chevalet C: **An assessment of European pig diversity using molecular markers: partitioning of diversity among breeds.** *Conserv Genet* 2005, **6**:729–741.
25. Rohrer GA, Freking BA, Nonneman D: **Single nucleotide polymorphisms for pig identification and parentage exclusion.** *Anim Genet* 2007, **38**:253-258.
26. Gu Z, Hillier L, Kwok PY: **Single nucleotide polymorphism hunting in cyberspace.** *Hum Mutat* 1998, **12**:221-225.
27. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: **Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms.** *Genome Res* 1998, **8**:748-754.
28. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Mining SNPs from EST databases.**

Genome Res 1999, **9**:167-174.
29. Kwok PY, Deng Q, Zakeri H, Taylor SL, Nickerson DA: **Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs.** *Genomics* 1996, **31**:123-126.
30. Garg K, Green P, Nickerson DA: **Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags.** *Genome Res* 1999, **9**:1087-1092.
31. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet* 1999, **23**:452-456.
32. Madsen BE, Villesen P, Wiuf C: **A periodic pattern of SNPs in the human genome.** *Genome Res* 2007, **17**:1414-1419.
33. Smith AFA, Green P: **RepeatMasker** [http://www.repeatmasker.org]
34. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7**:203-214.
35. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
37. Kent WJ: **BLAT—the BLAST-like alignment tool.** *Genome Res* 2002, **12:** 656-664.

3 Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey

Hindrik HD Kerstens[1], Richard PMA Crooijmans[1], Albertine Veenendaal[1], Bert W Dibbits[1], Thomas FC Chin-A-Woeng[2], Johan T den Dunnen[3], Martien AM Groenen[1]§

[1]Animal Breeding and Genomics Center, Wageningen University, Marijkeweg 40, Wageningen, 6709 PG, The Netherlands
[2]ServiceXS, Plesmanlaan 1d, Leiden, 2333 BZ, The Netherlands
[3]Leiden Genome Technology Center, Human and Clinical Genetics, Leiden University Medical Center, Einthovenweg 20, Leiden, 2333 ZC, The Netherlands
[§]Corresponding author

Email addresses:
HHDK: hindrik.kerstens@wur.nl
RPMAC: richard.crooijmans@wur.nl
AV: tineke.veenendaal@wur.nl
BWD: bert.dibbits@wur.nl
TFCC: t.chinawoeng@servicexs.com
JHD: ddunnen@humgen.nl
MAMG: martien.groenen@wur.nl

Published in BMC Genomics. 2009 Oct 16;10:479.

Abstract

Background

The development of second generation sequencing methods has enabled large scale DNA variation studies at moderate cost. For the high throughput discovery of single nucleotide polymorphisms (SNPs) in species lacking a sequenced reference genome, we set-up an analysis pipeline based on a short read de novo sequence assembler and a program designed to identify variation within short reads. To illustrate the potential of this technique, we present the results obtained with a randomly sheared, enzymatically generated, 2-3 kbp genome fraction of six pooled Meleagris gallopavo (turkey) individuals.

Results

A total of 100 million 36 bp reads were generated, representing approximately 5-6% (~62 Mbp) of the turkey genome, with an estimated sequence depth of 58. Reads consisting of bases called with less than 1% error probability were selected and assembled into contigs. Subsequently, high throughput discovery of nucleotide variation was performed using sequences with more than 90% reliability by using the assembled contigs that were 50 bp or longer as the reference sequence. We identified more than 7,500 SNPs with a high probability of representing true nucleotide variation in turkeys. Increasing the reference genome by adding publicly available turkey BAC-end sequences increased the number of SNPs to over 11,000. A comparison with the sequenced chicken genome indicated that the assembled turkey contigs were distributed uniformly across the turkey genome. Genotyping of a representative sample of 340 SNPs resulted in a SNP conversion rate of 95%. The correlation of the minor allele count (MAC) and observed minor allele frequency (MAF) for the validated SNPs was 0.69.

Conclusions

We provide an efficient and cost-effective approach for the identification of thousands of high quality SNPs in species currently lacking a sequenced genome and applied this to turkey. The methodology addresses a random fraction of the genome, resulting in an even distribution of SNPs across the targeted genome.

Background

The scalability and availability of highly automated genotyping assays for single nucleotide polymorphisms (SNPs) has made the SNP a popular marker in genetic linkage and association studies in a variety of species. In humans, large-scale identification and characterization has resulted in a repository of over 14 million SNPs [1] that are now being used in whole genome association studies to identify genes involved in complex genetic traits [2-6]. The availability of a high quality reference genome sequence and resources to perform low coverage resequencing on a few individuals are prerequisites for the traditional method of whole genome SNP discovery; genomic sequences of different individuals are aligned to a reference genome and nucleotide variation is detected [7]. Although very effective in species whose genome has been sequenced, such as human, cow, horse, and chicken, for the majority of species this method of SNP discovery is currently not feasible. Although second generation sequencing has lowered the cost per sequenced base a hundred-fold and allows the resequencing of complete genomes in a fraction of the time, the size of the sequencing target still exceeds the frequently available budget. By deep sequencing reduced representation libraries (RRL), SNPs can be discovered and allele frequencies estimated more economically [8]. The complexity of a pool of DNA samples from multiple individuals is reduced by two orders of magnitude [9] by isolating a fragment size range of a complete endonuclease digestion. Depending on the applied endonuclease, the obtained RRL contains hundreds of thousands of fragments within the optimum size range of the sequencing platform, equally distributed over the genome and with a low representation of repetitive elements. Tens of thousands of high quality SNPs can be identified by aligning the sequence reads that result from deep sequencing the RRL to a genome reference sequence. This approach already has been applied to species with a more or less completed genome draft sequence, like cow [8], as well as on species in which genome sequencing is ongoing, such as pig [10].

However, many species, such as turkey, are still lacking a completely sequenced genome. Although high-throughput sequencing technologies are rapidly evolving and have drastically lowered the cost of whole genome DNA sequencing, the de novo assembly of a mammalian-sized genome remains a challenge [11]. Despite the number of published algorithms for short fragment de novo sequence assembly [12-16], which assembles whole prokaryotic

genomes [17,18], reconstructing the sequencing targets of hundreds of megabases will require parallelization of these algorithms. Furthermore, many of these species still lack sufficient genetic markers and linkage maps that would aid in the ordering of the sequencing contigs and anchoring the contigs to specific chromosomes. Thus, the development of an efficient method for SNP discovery in such species is of high importance. We provide an effective strategy for combining RRL deep sequencing with de novo contig assembly based on next-generation sequencing data. The key of our approach is based on using RRLs consisting of large fragments (2-3 kbp) and random shearing. Performing high-throughput sequencing to a sufficient depth on sheared RRL in a pooled DNA sample in the first place enables reconstruction of the sampled genome fraction by de novo sequence assembly. The assembled contigs subsequently serve as a reference genome to which all short reads derived from multiple individuals can be mapped accurately, and SNPs can be called reliably [19].

The aim of this study was to develop an extremely cost effective method to detect high quality SNPs in unsequenced genomes. We applied this method to turkey, a species of considerable economic importance, and used the genome of a closely related species, chicken [20-22], to benchmark our approach.

Results

RRL preparation

We prepared a pooled DNA sample consisting of DNA samples from six turkey individuals. A RRL was prepared by digesting the pooled DNA sample with *Sau*3A and isolating the fragments in the size range of 2-3 kbp. This fraction consists of an estimated 5-6% of the turkey genome. The turkey genome has a high similarity to the chicken genome and is approximately the same size (~1.2 Gbp). Therefore, the isolated 5-6% fraction of the turkey genome represents approximately 62 Mbp. This estimate was confirmed by selecting all 2-3 kb fragments of an in silico *Sau*3A digest of the chicken genome build WASHUC2, which resulted in a total of 27,025 fragments representing 63.4 million bases. The turkey RRL was sequenced using the Illumina sequencing technique [23] after random shearing of the isolated *Sau*3A fragments. The resulting data set of short sequence reads forms the basis for contig assembly, providing sufficient sequence context flanking the SNPs to allow for the subsequent development of SNP genotyping assays.

DNA sequencing and sequence filtering

We generated 114 million sequence reads of turkey genomic DNA using the Illumina Genome Analyzer. The resulting 36 bp sequence reads were trimmed to 32bp because of the decay in base-call quality observed after the 32nd base. Subsequent removal of sequence reads with non-called bases resulted in almost 108 million reads, providing an estimated 56-fold coverage of 5-6% of the turkey genome. We used *Sau*3A to generate the RRL and, as expected, we observed that a fraction of the sheared DNA fragments started with the GATC restriction tag (Table 1), though the observed frequency was higher than expected.

Table 1: Summary of DNA sequence filtering results.

	Filter applied[1]	Pass Filter	(%)	GATC start (%)
Pre-selection	l32 n.	107888201	94.48	24.78
Assembly	l32 n. q20 o230	27979963	24.50	46.64
SNP	l32 n. q10 o230	32941906	28.84	40.23

[1] sequences are filtered for length 32, without base-call errors (n or .). Singly represented reads are required to have a per base-call quality score of 20 (assembly data set) or 10 (SNP data set). Sequences more than four times overrepresented, based on the expected 56X coverage, were discarded.

We discarded 984,258 reads tagged as repeat by RepeatMasker [24]. Reads that were, based on the theoretical coverage, over-represented more than four times were also removed because of their likeliness to resemble repetitive sequences or to represent duplicated regions in the turkey genome. Besides not being able to properly reconstruct repeats without mate-pair information at this low genome coverage, we also wanted to avoid false SNP predictions due to paralogous sequences. To improve the accuracy of the turkey genome assembly and reliably predict SNPs on the assembled contigs, data were screened for quality by applying a maximum sequencing error tolerance for reads with a single representation. For assembly purposes, we only tolerated one sequencing error per 100 bases, whereas one error per ten bases was tolerated in the reads used for SNP detection. After removing repetitive, overabundant, and low quality sequences with a single representation, almost 27 million reads (864 million bp) corresponding to 8.6 million unique sequences remained for contig

assembly. For SNP detection purposes, almost 33 million reads (1.05 billion bp) corresponding to 13.8 million unique sequences passed our thresholds.

Reference genome construction

For the actual SNP detection, a required reference genome was constructed by first performing de novo short read sequence assembly. Available de novo assemblers were SSAKE [12], SHARCGS [13], Edena [14], Velvet [15], and ALLPATHS [16]. Likely because of the large genome target and relatively high error rate of 1% ALLPATHS and SHARCGS showed an unfeasible large memory footprint and runtime. Probably because of the relatively low genome target coverage (14X) Velvet did return only 24 assembled sequence contigs all of which had a more than 15X coverage. Although Edena assembled contigs computationally more efficiently, SSAKE resulted in a higher number of assembled sequences and longer sequence contigs (Table 2).

Table 2: Short read assembly results.

	algorithm[1] and non-default parameters		
	edena -c 33 -m 16	velvet 15[2]	SSAKE
contigs	230741	24	627600
assembled reads	8965681	NA	13964267
assembly length	17487533	2812	36163074
N50	90	129	53

[1]algorithm versions were: edena-2.1.1, velvet_0.3, SSAKE_v2.02parameter applies to hash_length

Based on these results, the final assembly of the reference genome was performed with SSAKE. Using SSAKE [12], we assembled 36,163,074 bp into 627,600 short sequence contigs, with an average coverage of 9.52 and an N50 of 53 bp using the default assembly parameters. The quality of the reference genome assembly was estimated by mapping the short sequence contigs of at least 50 bp (further referred to as c50) to the draft genome sequence of chicken, the most closely related species [20-22] for which a genome sequence is available. As a benchmark, we used 20,000 publicly available turkey BAC-end sequences (BES) (Table 3). Direct alignment of the 35 bp Illumina reads resulted in the unique alignment of approximately one-third of the sequences. This fraction of turkey sequences uniquely aligning to the chicken genome steadily increased with increasing contig length, until reaching a maximum of

73% for contigs in the size range of 100-150 bp. At contig lengths above 150, this percentage gradually decreases dropping below 10% for contigs of 1000 bp and larger. The short sequence contigs and BES within the size range of 100-300 bp had comparable mapping statistics. The observed trend of a decrease in alignment for the larger assembled contigs was not observed for the BES.

Table 3: Quality estimation of turkey short read contigs based on alignment to the chicken genome.

	***frequency*[1]**		***percentage mapping to chicken*[2]**							
size	***contigs***	***BES***	***uniquely***		***with secondary hit***		***multiple hits***		***total***	
50-70	124480	*0*	47	-	2	-	1	-	50	-
75-100	38382	*1*	53	*0*	2	*0*	1	*0*	56	*0*
100-150	25808	*156*	73	*69*	1	*4*	0	*4*	74	*77*
150-200	8878	*226*	69	*78*	1	*6*	0	*4*	70	*88*
200-300	6453	*835*	63	*82*	1	*3*	0	*3*	64	*88*
300-400	2372	*2428*	51	*86*	1	*4*	0	*1*	52	*90*
400-500	1192	*6664*	40	*87*	1	*3*	0	*2*	40	*92*
500-600	682	*8509*	30	*88*	1	*3*	0	*2*	31	*93*
600-800	688	*1510*	18	*88*	0	*4*	0	*2*	18	*93*
800-1000	308	*54*	11	*83*	0	*0*	0	*2*	11	*85*
>=1000	380	*4*	6	*80*	1	*0*	0	*20*	7	*100*

[1] *frequency in which contigs and BES (in italics) occurred per size category*
[2] *per size category percentage of contigs and BES (in italics) that mapped to the chicken genome*

The distribution of the assembled contigs across the turkey genome was evaluated by aligning the contigs against the chicken genome. The contigs were distributed uniformly across the genome (Figure 1 and http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s1.png). The mapping results were subsequently used to further improve the assembly by merging contigs that mapped to adjacent or overlapping locations on the chicken genome. Merging of these contigs resulted in a more contiguous reference sequence and an increase in the average size of the assembled contig

(this assembly is further referred to as c50ca). We detected 15,754 adjacent or overlapping contigs, 13,695 identical overlaps, and 24,593 contigs in total were merged into 10,898 bigger contigs, representing 2,072,380 nucleotides and a N50 of 198 bp. Finally, we further extended our turkey reference genome (referred to as c50caB) by including the publicly available BES. A total of 5,831 BES (2,840,087 bp) with 49,638 short sequence contigs (4,032,887 bp) assembled into 8,526 new contigs with a total sequence length of 3,022,857 base pairs. The remaining 38,957,511 bp of the genome sequence was represented by 578,885 singletons. The BES, as well as all contigs from the extended assembly, were aligned to the chicken genome sequence by using BlastZ [25] to predict their distribution within the turkey genome (Figure 1 and http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s1.png).

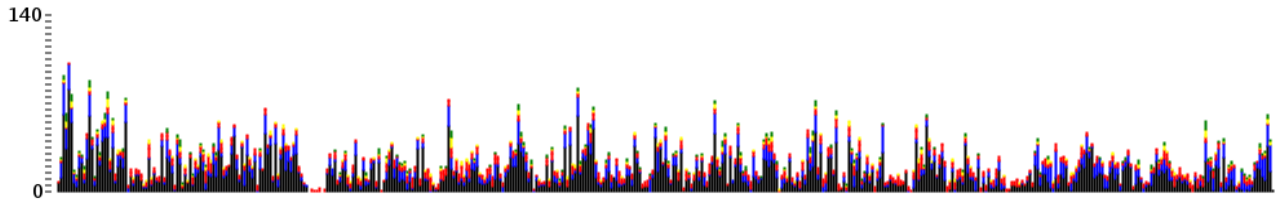

Figure 1: Distribution of short read turkey contigs, turkey BES, and SNPs on chicken chromosome 4. In black, short read contigs <100bp; in blue, short read contigs ≥100bp; in red, BES; in yellow, BES-short-read contigs; and in green, SNPs. On the X-axis, the chicken genome in 200 kb intervals. On the Y-axis, the frequency of mapped turkey features for a specific chicken genome interval.

SNP discovery

We aligned 32,941,906 reads (Table 1) to each of the three reference genomes described above (c50, 50ca, and c50caB). We adjusted the alignment parameters towards an approximately uniform distribution of nucleotide variation over the 32 bp reads using reference c50 (Figure 2). Putative SNPs within sequence clusters with a sequence depth less than four times the maximum theoretical coverage (58X), and in which the minor allele was represented at least three times, were recorded. Using these parameters, we identified 7,617 SNPs residing in 6,696 contigs out of the 209,623 contigs of the c50 reference. By using the C50ca assembly, 321 additional SNPs were detected (Table 4); furthermore, the fraction of SNPs with a sufficient flanking sequence increased considerably. Finally a further increase in the number of SNPs was achieved by using the reference assembly that included the BES (c50caB). This reference consisted of 192,731 contigs of which 7,952 contained

Figure 2: The X-axis represents the 32 base sequence read. On the Y-axis is the cumulative number of identified SNPs per base position of the sequence read.

one or more SNPs. Putative SNPs detected in uniquely mapped reference

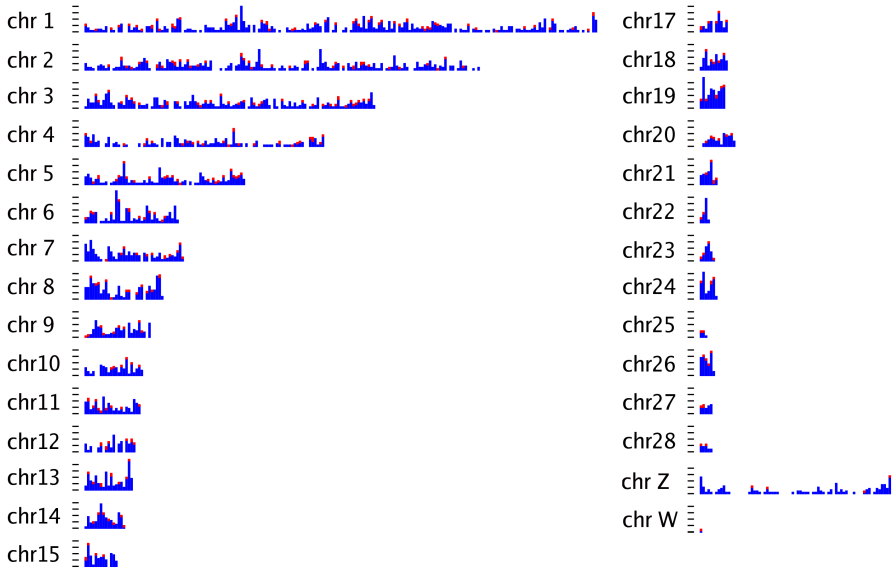

Figure 3. Distribution of 6,134 SNPs that mapped uniquely to the genome, 343 of which were selected for validation. In blue, 5,791 putative SNPs identified using the c50caB reference sequence and mapping uniquely to the chicken genome; in red, 343 uniquely mapping putative SNPs selected for validation. On the X-axis, the chicken genome in 1 Mb intervals. On the Y-axis, the frequency of mapped putative turkey SNPs for a specific chicken genome interval.

sequence contigs were plotted along the chicken chromosomes. Alignment with the chicken genome showed that the identified putative SNPs were distributed uniformly across the genome (Figure 3).

Validation

The application of the chicken reference genome in the improvement of our turkey reference, in which turkey contigs were merged based on comparative alignment results, requires conservation between these two genomes. Chicken and turkey genome conservation was determined by performing PCR amplification with forward and reverse primers designed on 13 neighboring short read turkey contigs aligning up to 0.5 kb apart on the chicken genome. As a control, PCR was performed on the corresponding chicken DNA target for which additional primer pairs were developed in the case that the turkey primers were not cross species applicable (http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s2.doc). The resulting PCR products on the turkey genome were compared with corresponding amplification products obtained on the chicken genome; they were approximately the expected length based on the chicken genome.

The contig assembly and SNP detection procedure were initially validated by PCR amplification and subsequent sequencing of the fragments in the six turkey individuals that made up the DNA pool from which the short read sequence data set was generated. Primers were developed on 12 contigs, each containing multiple putative SNPs. All 29 SNPs predicted on these 12 contigs were confirmed. In addition, a further five additional SNPs were identified (http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s2.doc**)**.

Further SNP validation was done by genotyping an animal panel consisting of 96 animals using 343 predicted SNPs distributed uniformly over the chicken genome (Figure 3) and 41 randomly selected SNPs that did not map uniquely to a single location in the chicken genome. A total of 340 SNPs gave reliable genotypes in the assay, and 96% of these were polymorphic (Table 5). We observed that SNPs predicted within contigs that uniquely mapped to the chicken genome had a more than five times higher chance of giving reliable genotypes than SNPs from contigs that aligned to multiple locations in the genome. The minor allele count (MAC) of each polymorphic SNP, the minor allele frequency (MAF) observed in the six animals represented in the discovery pool, and the MAF based on all 96 genotyped individuals are shown in

http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s3.xls. The average MAF of all successfully typed SNPs was 0.28, and the average heterozygosity in the individuals typed was 0.35. The correlation between MAC and MAF was 0.69 in the six animals that made up the discovery pool.

Table 4: Overview of SNPs identified.

			nt flanking sequence[1]		
reference[2]	**contigs**	**SNPs**	**40/40**	**20/20**	**2/40**
c50	209623	7609	2254	5218	6454
c50ca	195928	7930	2760	5636	6834
c50caB	192731	11287	5620	8902	10125

[1]40/40 refers to SNPs that are flanked on both sides by at least 40 nucleotides of genomic sequence.20/20 refers to SNPs that are flanked on both sides by at least 20 nucleotides of genomic sequence.2/40 refers to SNPs that have at least 2 nucleotides flanking sequence on one side and at least 40 nucleotides on the other.
[2]c50 refers to reference genome consisting of short read contigs of 50 bp or morec50ca is extended genome assembly based on chicken alignmentc50caB is extended genome assembly based on chicken alignment and turkey BES.

Table 5: SNP performance statistics.

SNP	**384 selected SNPs**			
	Mapped (343)	**%**	**Unmapped (41)**	**%**
performance				
polymorphic	304	88.6	20	48.8
monomorphic	12	3.5	4	9.8
not working	27	7.9	17	41.5

Genotyping performance of 343 SNP discovered in short read contigs that were uniquely mapped on the chicken genome and 41 SNPs discovered in contigs that were not, or not uniquely, mapped on the chicken genome.

Discussion

Next generation sequencing

Our large-scale nucleotide variation study on the turkey genome, including a partial assembly of a reference genome, demonstrates that short fragment second-generation sequencing of randomly sheared large fragment RRLs is an efficient and cost-effective approach for SNP discovery, providing thousands of high quality SNPs, even in the absence of an available genome sequence. This approach combines the advantages of using an extremely cost-effective sequencing platform with the ability to provide SNP sequence context by short fragment assembly. The sequence context provided by this SNP detection approach makes this the ideal method for the development of SNP assays on a variety of genotyping platforms for all species without sequenced genomes.
We had to discard nearly 75% of our sequence data to meet quality constraints for the sequence assembly (Table 1). This was in pair due to suboptimal sequence densities resulting in suboptimal clustering on the tiles of the Illumina Genome Analyzer (see methods section), resulting in poor sequence quality and low sequence output. On top of this, a relatively large proportion (about half of the sequences passing our quality thresholds) started with the endonuclease cleavage site. The underestimation of this fraction in the initial length trimmed sequencing data subset was most likely caused by sequencing errors in the first four bases of a read. Stringent filtering of reads revealed the real ratio and provided a higher quality data subset, but lowered the total theoretical coverage of our sequencing target to 10X. To avoid this observed bias towards the ends of the RRL fragments, an option is to dephosphorylate the ends of the RRL restriction fragments prior to random shearing and ligation to the sequencing adaptors. We were only able to assemble roughly 60% of our sequencing target covered by our RRLs, most likely due to the limited sequence depth (10X) of our final data set after using stringent quality thresholds. The recent addition of paired end sequencing to second generation sequencing, the increased read length and the predicted further increase in sequence length and tens of gigabases of useful sequence data per machine run in the near future [23,26], will allow more efficient sequence assembly. This will result in increased coverage of the sequence target and an increased contig length of the assembled sequences, at lower costs. An improved assembly allows a substantial increase in the number of SNPs identified, as well as a considerable increase in the number of SNPs for which a genotyping assay can be designed. Another strategy to increase the number of assayable SNPs would be to use combination

of two different sequencing platforms, such as Roche 454 and Illumina GA, in which longer reads (454) are being used for reference construction and short reads provide the necessary sequencing depth to detect nucleotide variation.

Benchmarking and improving

We showed that the genome sequence of a closely related species can be used for benchmarking the assembled contigs, the genome coverage and can further improve the reference assembly by merging contigs mapping to an adjacent location on the genome of that particular species. In the case of turkey, we applied this cost-effective strategy by using the likewise Galliform genome of the chicken. Previous studies indicate that chicken and turkey karyotypes (common ancestor ~28 MYA) have undergone relatively very few chromosomal rearrangements during evolution compared to mammals [20]. Moreover, results of cross species hybridisation studies and comparative genomics suggest that chicken an turkey share a high sequence identity [20-22] which makes the chicken genome sequence usable to benchmark the turkey reference assembly.
An assessment of the quality of our assembled turkey contigs was done by mapping the contigs to the chicken genome and comparing the results with the alignment statistics of turkey BES of the same size range. The results indicate that the contigs up to 300 bp, in general, are of good quality and that turkey BES share high sequence identity with the chicken genome. The comparison between the assembled contigs and BES indicate that most of these contigs represent valid sequences of the turkey genome. At increasing contig length, the number of sequences that align uninterrupted to a unique location in the chicken genome declines, dropping below 10% for contigs in the size range of 1000 bp. The fact that this decline is not observed for the turkey BES indicates that it is not due to small indels between the chicken and turkey sequences, but that this is an artifact caused by the assembly. These results indicate that at increasing contig lengths, the chance of mis-assemblage by SSAKE increases exponentially. However, because most SNP typing assays make use of the sequences directly flanking the SNP, this will only have a small effect on the success rate of the genotyping assays. At total of 7609 SNPs were identified on the assembled short read contigs of which 84% was flanked by sufficient sequence to allow probe design in a genotyping assay. To make the turkey reference more contiguous we used the chicken genome to identify contig pairs that uniquely mapped adjacent to each other, showing a small overlap. In 87%

of these cases, overlapping contigs appeared to have identical sequences within the overlapping region. Although biased by the alignment algorithms used, which remove unaligned tailing ends of contigs, our comparative assembly results suggest that the mapped contigs are of a constant high quality and can be mapped with high accuracy. Therefore, these results allow the merging of the smaller contigs, resulting in a significant increase in the average length of the assembled turkey contigs. The resulting reference sequence appeared to be beneficial in the identification of SNPs and, in particular, increased the number of SNPs with sufficient flanking sequence for designing a genotyping assay. This benefit is clearly illustrated by the 4% increase in the total number of SNPs identified and 22% increase in SNPs with at least 40 bp of flanking sequence on both sides. The alignment of the turkey contigs with the highly similar chicken genome also turned out to be a good predictor of genotyping success rates for the SNPs (Table 5). The SNPs located on turkey contigs that aligned to more than a single location on the chicken genome appeared more likely to fail in the genotyping assay than SNPs located on uniquely aligning turkey contigs which is probably because these are likely to contain duplicated sequence or repetitive sequences of the turkey genome. Repetitiveness of turkey and chicken genome sequences were compared by applying the IR [27] algorithm on the available turkey BES (9,9 Mb) and 20,000 (8,3Mb) chicken genomic sequences randomly selected from the NCBI database (data not shown). Obtained non-normalised I_r values suggest that the turkey genome is slightly less repetitive (0.6247) than the chicken genome (0.7126). The average I_r for the chicken genome was 0.3905 and ranged from 0.0793 in chromosome 19 to 1.3419 in chromosome 16. Compared to other eukaryotes like Human, Mouse and Arabidopsis [27] the chicken genome is at least three times less repetitive which is in line with the results of a previous study in which repeats were computationally identified on the chicken genome [28]. This lower level of repetitiveness is beneficial for the genotyping success rate because of the lower occurrence of false SNP predictions due to repetitive genomic regions.

To further maximize the number of identified SNPs, the available turkey BES were added to the reference genome. Again, these additional sequences not only resulted in the identification of an additional 3357 additional SNPs, they also increased the number of SNPs with a sufficient amount of flanking sequence. The assembly of short read contigs and BES resulted in, at least, a 25% reduction of sequence redundancy in the assembled short read contigs. Removal of sequence redundancy in the reference genome is beneficial for downstream SNP detection because of the reduction in the number of sequence reads being

assigned ambiguously to multiple locations on the reference genome during the alignment. SNPs predicted within sequence clusters containing these ambiguously mapped reads are indistinguishable from falsely predicted SNPs due to the clustering of paralogous sequences and thus discarded.

Allele frequencies

Our conservative approach requiring a minimal MAC of three was designed to minimize false positive SNP discovery and, consequently, ignored large numbers of less abundant true nucleotide variations. The five additional SNPs we identified by PCR and sequencing that were not previously detected in silico are a typical consequence of applying a minimum redundancy cut-off. However, the selection for SNPs with a MAC of at least three drastically reduces the chance that sequencing errors are considered an SNP. Keeping the number of false positives as low as possible in general is more important than maximizing the number of SNPs. True nucleotide variation might also be lost during sequence assembly in which contigs are extended by a read only if the consensus base ratio is 0.6 or more. Single nucleotide polymorphisms with a MAF higher than 0.4 very likely break the contig extension; for this reason, they will only be detected on the tailing ends of assembled contigs. The absence of sequence context on one side of these polymorphisms further hampers the alignment of additional reads to form deep sequence clusters meeting the minimum allele count constraint applied during SNP detection. This concept explains the increase in the number of SNPs discovered on the extended reference genome though the number and total number of base pairs covered decreased. The occurrence of a few SNPs with an estimated MAC higher than 0.4 can be explained by a lower MAC in the assembly data subset compared to the MAC in the SNP detection data subset.

Conclusions

Our strategy of assembling a reference genome from short next-generation sequences of a randomly sheared RRL of pooled genomes, followed by subsequent SNP detection by aligning the same short reads against this reference genome, is a cost-effective and efficient method for the high rate discovery of SNPs in species with unsequenced genomes. The availability of a closely related sequenced genome is not a requirement but comparative mapping facilitates the selection for high quality SNPs. Our comparison with

the chicken genome further suggests that the high quality SNPs identified in this report most likely cover the complete turkey genome and provide the first large SNP resource for genetic studies in turkey.

Methods

Library construction

Genomic DNA was extracted from the blood of six unrelated F_0 individuals from a male and a female turkey line, selected for growth and reproduction characteristics respectively, three samples from each line. The selection of the restriction enzyme was based on the 10 to 20-fold reduction of genome complexity in the 2-3 kb size region run on a 1.5% agarose gel. Ten enzymes were tested (*Sau3A, XhoI, AvaI, MspI, SacI, KpnI, SalI, AluI, TagI*; New England Biolabs, Ipswich, MA, USA); of which, *Sau3A* was finally selected to make the Turkey RRL because of good digestion performance and a desired 5-6% fraction of the genome in the 2-3KB size range. In total, 100 µg of the pooled DNA was digested using 1,000 units of the restriction enzyme *Sau3*A in a total volume of 240 µl. The digested pooled DNA sample was fractionated on 1.5% low melting point agarose gel at 100V for 3 hours and stained with ethidium bromide. The 2-3 kb sized fraction was sliced out of the gel, melted, and loaded on a new 1.5% low melting agarose gel for another fractionation at 100V for 1 hour. The 2-3 kb fraction was sliced out of the gel and the DNA was recovered by ß-Agarase-I treatment, purified by phenol-chloroform extraction, and precipitated with 2-propanol. DNA was dissolved in TE with a concentration of 50 ng/µl. The isolated DNA was randomly sheared, end-repaired, and prepared using the Illumina Sample preparation kit [29].

Sequencing

Five picomole aliquots of the library were processed with the Illumina Cluster Generation Station (Illumina Inc., USA) following the manufacturer's recommendations. The Illumina IG Genome analyzer (Illumina Inc., USA) was programmed to produce a theoretical fixed read length of 36 bp. Images were collected over 4,040 tiles, each of which contained 685-41,954 clusters.

Sequence filtering and reference assembly

Reads were trimmed to 32 bp, and reads with an occurrence of more than four times the theoretical coverage were discarded. Two data sets were created; one was the assembly data set and the other the SNP detection data set. In the SNP

detection data set, we required a per base quality score of at least 10 if the read was singly represented. For the assembly data set, we required that a particular 32 bp sequence be represented two times or that every base in the 32 bp sequence have a quality score of at least 20.
Furthermore, the assembly data set was analyzed for repetitive elements using RepeatMasker [24] with default options, species chicken, and reads containing repetitive elements were removed. Remaining reads were assembled to short read contigs using SSAKE [12] and the default parameters. The data set containing contigs larger than 50 bp are referred to reference genome c50.
The short read contigs (c50) were mapped on the chicken genome with the selection criteria that a contig had to align along 80% of its length with at least 60% identity. Short read contigs in the size range of 50-100 bp were mapped using Megablast [30], and short read contigs of 100 bp and longer were mapped using BlastZ [25]. Mapping results were parsed using a custom made Perl script to identify short read contigs that mapped adjacent or with a less than 21 bp identical overlap. These identified contigs were subsequently merged, and this data set is referred to as reference genome c50ca.
The turkey genome reference sequence was further extended by adding 20,388 publicly available BES of the CHORI-260 turkey BAC library [31] to all short read contigs (data set c50ca) and assembled using phrap [32] and the default parameters. Obtained sequences larger than 50 bp were used as a turkey reference genome in the SNP detection procedure and referred to as c50caB.

SNP detection

The SNP detection was performed with MAQ [19] (default parameters) using the SNP detection data set and one of the reference genomes (c50, c50ca, or c50caB). Putative SNPs were tagged if the reads involved were mapped unambiguously on the reference genome and the minor allele appeared at least three times. The SNPs were discarded if the depth exceeded four times the theoretical sequence depth, the consensus quality of the SNP was less than 30, or the best mapping read in the sequence cluster had a mapping score lower than 60.

Validation

Validation of the assembled contigs and detected SNPs was performed two ways.

First, PCR primers were designed for 12 contigs containing multiple SNPs using primer 3. The PCR was performed in 12 μl and contained 6 μl Abgene 2x PCR Mastermix (ThermoScientific), 60 ng template DNA, and 4 pmol of each of the two primers. The PCR cycling conditions were 95°C for 5 min, 35 cycles of 30 s at 95°C, 45 s at 55°C, and 90 s at 72°C, followed by a final elongation step of 72°C for 2 min.

The PCR products of the six animals from the discovery panel were purified using millipore PCR cleanup filter plates (MSNU03050) and sequenced using the DETT sequencing kit according to the manufacturer's specifications (GE Healthcare). Unincorporated dye terminator was removed by ethanol precipitation and analyzed on a 48-capillary ABI 3730 DNA analyzer (Applied Biosystems). Sequencing results were further analyzed with the STADEN Package.

The second method of validation was genotyping the SNPs using the Illumina GoldenGate® Genotyping assay on an Illumina® BeadXpress with veraCode™ technology. Selection criteria for the SNPs were based on the Illumine design score (above 0.8) and MAC ranging from .5 to .15 detected by MAQ [19]. For the total 384 SNPs assayed, including 343 SNPs equally distributed along the chicken genome and 41 randomly selected SNPs that did not map to a single location in the chicken genome, oligonucleotides were designed, synthesized, and assembled into oligo pooled assays (OPA) by Illumina Inc. The 384 SNPs were genotyped in 96 animals which included the six F_0 animals from the discovery panel and 29 additional F_0 animals and further consisted of 47 F_1 animals and 14 unrelated animals derived from 2 inbred lines. Genotyping results were analyzed in Beadstudio.

The correlation between allele frequency estimated by sequencing and genotyping was calculated over 310 observations (http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s3.xls) by randomly selecting the major or minor allele.

Availability and requirements

The SNPs identified in this study, in which the polymorphism is flanked by 20 bp of sequence context on each side, have been deposited in the National Center of Biotechnology (NCBI) SNP database (dbSNP) under submitter handle WU_ABGC. NCBI_ss 142460378-142468928 excluding (142463311, 142463314, 142463316, 142463318, 142463320, 142463322, 142463324, 142463326, 142463328, 142463330, 142463332, 142466905, 142466907, 142466910, 142466912) represents predicted SNPs that were not tested in our

animal panel. Predicted SNPs that were confirmed are listed in http://www.biomedcentral.com/content/supplementary/1471-2164-10-479-s3.xls. The SNPs with less than a 20 bp sequence context will be available upon request.

Authors' contributions

HHDK designed and developed the genome reference construction method and SNP prediction method and wrote the manuscript. AV, BWD, and RPMAC collected and prepared the samples and performed the initial validation and genotyping analysis. TFCC, JHD, and RPMAC coordinated and supervised the DNA sequencing. MAMG and RPMAC coordinated and supervised the experiment implementation and assisted in the manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

We thank Mari Smits, Nikkie van Bers, Robert Kraus, and Hendrik-Jan Megens for critically reading the manuscript and their helpful comments. This study was funded by European Union grant FOOD-CT-2004-506416 (Eadgene), Netherlands National Computing Facilities foundation grant SH-018-07, and Hendrix Genetics, the Netherlands.

References

1. Sherry ST, Ward M, Sirotkin K: **dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.** *Genome Res* 1999, **9**: 677-679
2. Trikka D, Fang Z, Renwick A, Jones SH, Chakraborty R, Kimmel M, Nelson DL: **Complex SNP-based haplotypes in three human helicases: implications for cancer association studies.** *Genome Res* 2002, **12**: 627-639
3. Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, Kirby A, Daly MJ, Jager PLD, Walsh E, Lander ES, Rioux JD, Hafler DA, Ivinson A, Rimmler J, Gregory SG, Schmidt S, Pericak-Vance MA, Akesson E, Hillert J, Datta P, Oturai A, Ryder LP, Harbo HF, Spurkland A, Myhr K, Laaksonen M, Booth D, Heard R, Stewart G, Lincoln R, Barcellos LF, Hauser SL, Oksenberg JR, Kenealy SJ, Haines JL, Consortium IMSG: **A high-density screen for linkage in multiple sclerosis.** *Am J Hum Genet* 2005, **77:** 454-467
4. Consortium WTCC, (TASC) ASC, Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI,

Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans DM, Leung H, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, Clair DS, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskivina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop TD, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Matthew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop MG, Connell J, Dominiczak A, Marcano CAB, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Genetics BiR, Committee GSS(S, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hilder SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DPM, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JRB, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Hill AVS, Bradbury LA, Farrar C, Pointon JJ, Wordsworth P, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Gough SCL, Seal S, (UK) BCSC, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Newport M, Sirugo G, Rockett KA, Bumpstead SJ, Chaney A, Downes K, Ghori MJR, Gwilliam R, Hunt SE, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Widden C, Withers D, Cardin NJ, Davison D, Ferreira T, Pereira-Gale J, Hallgrimsdo'ttir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Brown MA, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AVS, Parkes M, Pembrey M, Stratton MR, Mitchell SL, Newby PR, Brand OJ, Carr-Smith J, Pearce SHS, McGinnis R, Keniry A, Deloukas P, Reveille JD, Zhou X, Sims A, Dowling A, Taylor J, Doan T, Davis JC, Savage L, Ward MM, Learch TL, Weisman MH, Brown M: **Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants.** *Nat Genet* 2007, **39**: 1329-1337

5. Meyre D, Delplanque J, Chèvre J, Lecoeur C, Lobbens S, Gallina S, Durand E, Vatin V, Degraeve F, Proença C, Gaget S, Körner A, Kovacs P, Kiess W, Tichet J, Marre M, Hartikainen A, Horber F, Potoczna N, Hercberg S, Levy-Marchal C, Pattou F, Heude B, Tauber M, McCarthy MI, Blakemore AIF, Montpetit A,

Polychronakos C, Weill J, Coin LJM, Asher J, Elliott P, Järvelin M, Visvikis-Siest S, Balkau B, Sladek R, Balding D, Walley A, Dina C, Froguel P: **Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations.** *Nat Genet* 2009, **41**: 157-159

6. Rafnar T, Sulem P, Stacey SN, Geller F, Gudmundsson J, Sigurdsson A, Jakobsdottir M, Helgadottir H, Thorlacius S, Aben KKH, Blöndal T, Thorgeirsson TE, Thorleifsson G, Kristjansson K, Thorisdottir K, Ragnarsson R, Sigurgeirsson B, Skuladottir H, Gudbjartsson T, Isaksson HJ, Einarsson GV, Benediktsdottir KR, Agnarsson BA, Olafsson K, Salvarsdottir A, Bjarnason H, Asgeirsdottir M, Kristinsson KT, Matthiasdottir S, Sveinsdottir SG, Polidoro S, Höiom V, Botella-Estrada R, Hemminki K, Rudnai P, Bishop DT, Campagna M, Kellen E, Zeegers MP, Verdier Pd, Ferrer A, Isla D, Vidal MJ, Andres R, Saez B, Juberias P, Banzo J, Navarrete S, Tres A, Kan D, Lindblom A, Gurzau E, Koppova K, Vegt Fd, Schalken JA, Heijden HFMvd, Smit HJ, Termeer RA, Oosterwijk E, Hooij Ov, Nagore E, Porru S, Steineck G, Hansson J, Buntinx F, Catalona WJ, Matullo G, Vineis P, Kiltie AE, Mayordomo JI, Kumar R, Kiemeney LA, Frigge ML, Jonsson T, Saemundsson H, Barkardottir RB, Jonsson E, Jonsson S, Olafsson JH, Gulcher JR, Masson G, Gudbjartsson DF, Kong A, Thorsteinsdottir U, Stefansson K: **Sequence variants at the TERT-CLPTM1L locus associate with many cancer types.** *Nat Genet* 2009, **41**: 221-227

7. Li G, Ma L, Song C, Yang Z, Wang X, Huang H, Li Y, Li R, Zhang X, Yang H, Wang J, Wang J: **The YH database: the first Asian diploid genome database.** *Nucleic Acids Res* 2009, **37**: D1025-D1028

8. van Tassell CPV, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS: **SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.** *Nat Methods* 2008, **5**: 247-252

9. Altshuler D, Pollara VJ, Cowles CR, Etten WJV, Baldwin J, Linton L, Lander ES: **An SNP map of the human genome generated by reduced representation shotgun sequencing.** *Nature* 2000, **407**: 513-516

10. Wiedmann RT, Smith TPL, Nonneman DJ: **SNP discovery in swine by reduced representation and high throughput pyrosequencing.** *BMC Genet* 2008, **9:** 81

11. Holt RA, Jones SJM: **The new paradigm of flow cell sequencing.** *Genome Res* 2008, **18**: 839-846

12. Warren RL, Sutton GG, Jones SJM, Holt RA: **Assembling millions of short DNA sequences using SSAKE.** *Bioinformatics* 2007, **23**: 500-501

13. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.** *Genome Res* 2007, **17**: 1697-1706
14. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J: **De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18**: 802-809
15. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008 **18**: 821-829
16. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: de novo assembly of whole-genome shotgun microreads.** *Genome Res* 2008, **18**: 810-820
17. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes.** *Genome Res* 2008, **18**: 324-330
18. Farrer RA, Kemen E, Jones JDG, Studholme DJ: **De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads.** *FEMS Microbiol Lett* 2009, **291**: 103-111
19. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**: 1851-1858
20. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RPMA, Groenen MAM, Deryusheva S, Gaginskaya E, Carré W, Waddington D, Talbot R, Völker M, Masabanda JS, Burt DW: **Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution.** *BMC Genomics* 2008, **9**: 168
21. Reed KM, Faile GM, Kreuth SB, Chaves LD, Sullivan LM: **Association and in silico assignment of sequences from turkey BACs.** *Anim Biotechnol* 2008, **19**: 80-83
22. Chaves LD, Knutson TP, Krueth SB, Reed KM: **Using the chicken genome sequence in the development and mapping of genetic markers in the turkey (Meleagris gallopavo). Anim Genet** 2006, **37**: 130-138
23. Illumina: [http://www.illumina.com/
24. Smith AFA, Green P: **RepeatMasker** [http://www.repeatmasker.org]
25. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: Human-mouse alignments with BLASTZ. Genome Res 2003, 13: 103-107
26. Applied Biosystems: [http://www.appliedbiosystems.com/]
27. Haubold B, Wiehe T: **How repetitive are genomes?** *BMC Bioinformatics* 2006, **7**: 541
28. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**: 695-716

29. Illumina: **Protocol for Whole Genome Sequencing using Solexa Technology.** *BioTechniques Protocol Guide* 2006, **12**: 29
30. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7**: 203-214
31. Nenfedov M, Zhu B, Thorsen J, Shu CL, Cao Q, Osoegawa K, Jong Pd: **New chicken, turkey, salmon, bovine, porcine and sheep genomic BAC libraries to complement world wide effort to map farm animals genomes.** P*lant and Animal Genome XI Conference Scherago International* 2003, 96, Abstract P87
32. Green P: **Phrap** [http://www.phrap.org]

4 Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries

Hindrik HD Kerstens[1,] Richard PMA Crooijmans[1], Bert W Dibbits[1], Addie Vereijken[2], Ron Okimoto[3] and Martien AM Groenen[1§]

[1]Animal Breeding and Genomics Center, Wageningen University, Marijkeweg 40, 6709 PG, Wageningen, the Netherlands
[2]Research and Technology Centre, Hendrix Genetics, P.O. Box 30, 5830 AE, Boxmeer, The Netherlands
[3]Cobb-Vantress Inc, P.O. Box 1030, Siloam Springs, AR 72761, USA
[§]Corresponding author

Email addresses:
HHDK: Hindrik.Kerstens@wur.nl
RPMAC: Richard.Crooijmans@wur.nl
BWD: Bert.Dibbits@wur.nl
AV: Addie.Vereijken@hendrix-genetics.com
RO: Ron.Okimoto@cobb-vantress.com
MAMG: Martien.Groenen@wur.nl

Submitted

Abstract

Background

Variation within individual genomes ranges from single nucleotide polymorphisms (SNPs) to kilobase, and even megabase, sized structural variants (SVs), such as deletions, insertions, inversions, and more complex rearrangements. Although much is known about the extent of SVs in humans and mice, species in which they exert significant effects on phenotypes, very little is known about the extent of SVs in the 2.5-times smaller and less repetitive genome of the chicken.

Results

We identified hundreds of shared and divergent SVs in four commercial chicken lines relative to the reference chicken genome. The majority of SVs were found in intronic and intergenic regions, and we also found SVs in the coding regions. To identify the SVs, we combined high-throughput short read paired-end sequencing of genomic reduced representation libraries (RRLs) of pooled samples from 25 individuals and computational mapping of DNA sequences from a reference genome.

Conclusion

We provide a first glimpse of the high abundance of small structural genomic variations in the chicken. Extrapolating our results, we estimate that there are thousands of rearrangements in the chicken genome, the majority of which are located in non-coding regions. We observed that structural variation contributes to genetic differentiation among current domesticated chicken breeds and the Red Jungle Fowl. We expect that, because of their high abundance, SVs might explain phenotypic differences and play a role in the evolution of the chicken genome. Finally, our study exemplifies an efficient and cost-effective approach for identifying structural variation in sequenced genomes.

Background

Structural variation within the genome, including insertions, duplications, deletions, and inversions of up to multiple kilobase pairs, have recently been described in a variety of species, including humans [1-3], mice [4], rats [5], silkworms [6] drosophila [7], and dogs [8]. These genomic variations were recently found to be widespread, encompassing 5% of the human genome [9],

and are thought to be involved in (co)determining complex phenotypes [10,11]. The contribution of structural variants (SVs) to complex phenotypes has been measured by association analyses of variance in gene expression levels (traits) and the presence of SVs. SNPs and SVs have been shown to account for 83.6% and 17.7%, respectively, of the total detected genetic variation in gene expression, with only a limited overlap [12]. The effect that SVs have on gene expression is likely underestimated given the much less completeness and accuracy with which SVs could be queried at that time. In humans, SVs have been associated with sporadic and Mendelian diseases, such as Williams-Beuren syndrome, mental retardation, and red-green color blindness. SVs have also been associated with complex human traits, such as autism, schizophrenia, Crohn's disease, and susceptibility to HIV infection [13]. Because of their association with human diseases, the importance of SVs has become increasingly apparent [9,14,15]. For most other species, including the major farm animals, chickens, cattle, and pigs, the extent and biological consequences of SVs have remained largely unknown due to the lack of a cost-effective approach for detecting SVs.

Until recently, comparative genomic hybridization (array-CGH) was the most commonly used method for detecting SVs [16]. Fosmid paired-end sequencing, which is a more laborious technique, has been used to detect SVs larger than 8 kb [17,18]. The inability to resolve smaller SVs using array-CGH results in the over-representation of larger SVs in current databases of structural variation (e.g., http://projects.tcag.ca/variation/). The resolution of array-CGH, though extremely costly, can be improved by using high-resolution whole-genome tiling arrays. Most of these SVs have been identified by methods that do not resolve SV end points at the base pair level. In addition, methods like array-CGH are based on a reference genome that currently does not encompass all SVs within the population and, thus, is limited in scope. Genomic regions that are the result of deletions not present in the reference genome are not captured by the array and not analyzed for SVs.

Next generation sequencing (NGS) technology was recently shown to be a powerful alternative to array-CGH for identifying genomic structural variation [1,7,19]. Using paired-end sequencing, SVs can be identified with single base pair resolution, and balanced rearrangements in which there is no gain or loss of a genomic region, such as inversions and translocations, cannot be identified by array-CGH. Paired-end sequencing and mapping (PEM) involves sequencing

the paired ends of fragments of known insert size from a genomic DNA library and computationally mapping DNA reads to a reference genome.
Here, we used PEM on reduced representation libraries (RRLs) of pooled chicken DNA samples.
In the chicken genome, only 43 (larger) SVs have been described thus far [20]. These SVs encompass 16 chicken-turkey inter-specific copy number variants (CNV) and 32 chicken-duck inter-specific CNVs, of which five CNVs overlap with inter-specific chicken-turkey CNVs [21]. In chicken, some phenotypes have already been linked to structural variation, including the pea-comb [22] and late feathering [23] phenotypes. With PEM of an RRL, we provide a cost-effective approach for exploring the presence of SVs at high resolution within four chicken breeds.

Results

Paired-end sequencing and mapping

To identify genomic rearrangements in the chicken genome, we applied massively parallel sequencing using the Illumina Genome Analyzer platform to sequence both ends of the genomic DNA fragments derived from the RRLs. We used pooled samples from 25 individuals to construct *Alu*I RRLs for a white egg layer line, brown egg layer line, and two different broiler lines. For the white and brown egg layer lines, the 150-200 bp *Alu*I fragments were used for creating the RRL; for the two broiler lines, 125-200 bp *Alu*I fragments were used. From the brown and white egg layer RRLs, we obtained 31.61 million and 29.70 million raw reads, respectively, and from broiler 1 and broiler 2 we obtained a total of 34.8 million and 32.4 million raw reads, respectively. Reads were filtered for the presence of the restriction enzyme tag and trimmed to 32 bases. We required a phred quality score [24] of at least 20 (Table 1) for each base in the 32-bp read. The fraction of read pairs for which both reads mapped back to the reference chicken genome (Red Jungle Fowl built WASHUC2) was 55% for broiler 1 and 70% for broiler 2 (Table 1). In the layers, the fraction was 66% (brown egg layer) and 62% (white egg layer). Of the approximately hundred thousand paired reads in all breeds, only one read (0.4-0.5%) mapped back to the reference genome, whereas up to 44% of the read pairs had no end mapping back to the reference genome.

Table 1: Sequencing and mapping results for the four chicken breeds analyzed for structural variation.

	Sequencing		*Mapping*						
Breed	***Raw reads***	***Paired l32q20[1]***	***Concor-dant[2] %***	***Neither end[3] %***	***One end[4] %***	***Diff chr[5] %***	***Too short[6]***	***Too long[7]***	***Relative orientation[8]***
Brown egglayer	31.61	23.59	66.26	33.17	0.45	0.02	417	19252	480
White egglayer	29.70	21.84	61.95	37.15	0.54	0.12	894	18836	1862
Broiler 1	34.82	24.83	70.40	29.06	0.43	0.01	1837	19086	303
Broiler 2	32.28	20.64	55.33	44.08	0.37	0.04	4458	16525	724

Paired-end sequencing of RRLs resulted in the indicated number of raw reads per breed. Sequencing read counts are in millions. Mapping percentages are relative to Paired l32q20.1Paired l32q20 = paired reads had the RRL restriction tag trimmed to 32 bp and were filtered for a minimum per base quality of 20; 2Concordant = both reads of a read pair mapped to the expected orientation relative to each other and in the expected distance according to the RRL size range; 3Neither end = none of the reads of a read pair mapped to the reference; 4One end = only one read of a read pair was mapped; 5Diff chr = both reads of a read pair mapped, but to different chromosomes; 6Too short = both reads of a read pair mapped to the expected orientation relative to each other but at a closer distance than expected based on the RRL size range; 7Too long = both reads of a read pair mapped at a larger distance from each other than expected; 8Relative orientation = reads of a read pair mapped in another orientation relative to each other than expected based on the reference chicken genome.

To calculate the sequence coverage of the RRL, we estimated the number of fragments in the RRL by performing an *in silico Alu*I digest of the chicken genome build WASHUC2, which resulted in 583,826 fragments of 150-200 bp, whereas 947,538 fragments of 125-200 bp were obtained. We calculated RRL sequence coverage based on the paired-end reads that passed our sequence quality filters. Coverage of the RRLs ranged from 11-13X in broiler lines to 18-20X in the layer lines, indicating that we analysed, on average, 22-40% of the haplotypes of the 25 individuals used for constructing the RRL (Table 2).
For each breed, we calculated insert sizes for paired ends that mapped in the

correct orientation (Figure 1). The results show a peak at ~185 bp and a shoulder of smaller fragments, indicating that the insert sizes were not equally distributed. The upper limit of fragment size was clearly demarcated at ~210 bp, which corresponded well to the size range of the excised fragments. Based on these results, the lower limit was estimated to be ~135 bp in the layer lines and ~110 bp in the broiler lines, which is consistent with the applied size selection. To eliminate false positives, we established size thresholds of 100 and 220 bp and considered mapping paired reads within this range as consistent with the reference genome.

Table 2: RRL construction simulated by an in silico AluI digest of the WASHUC2 build of the reference chicken genome.

Line	*Size-range*	*Number of fragments*	*Genome fraction*	*Sequenced (32 bp reads)*	*RRL coverage*
Layers	150-200	583826	101 Mb (8%)	18.7 Mb (1.5%)	37-40X
Broilers	125-200	947538	151 Mb (12%)	30.3 Mb (2.4%)	22-26X

Fragments were collected in corresponding size ranges as used in the in vitro RRL preparation. The total number of collected fragments and number of bases captured are indicators of what genome fraction was sampled. Based on trimmed reads, the fraction of the genome actually sequenced was calculated. The number of raw read pairs obtained (see Table 1) divided by the number of fragments is an indicator of the RRL coverage.

Rearrangements

In each breed, roughly 0.1% of the mapping read pairs had no concordant alignment in the reference genome, referred to as discordant paired-end reads [2,17], indicating a potential SV. Discordantly mapping read pairs are pairs that map too short or too far of a distance in base pairs according to the RLL size range or in another relative orientation than expected based on the reference genome (Table 1). Paired reads that mapped to two different chromosomes (up to 0.12%) were excluded from further analysis. Discordantly mapping read pairs with similar mapping coordinates and predicting a similar putative SV were clustered in 10,559 clusters. Clusters were classified as having an insert size that was too large (deletions, n=5135), too small (insertions, n=5241), or an incorrect orientation of ends (inversion breakpoints, n=183) with respect to the chicken genome sequence.

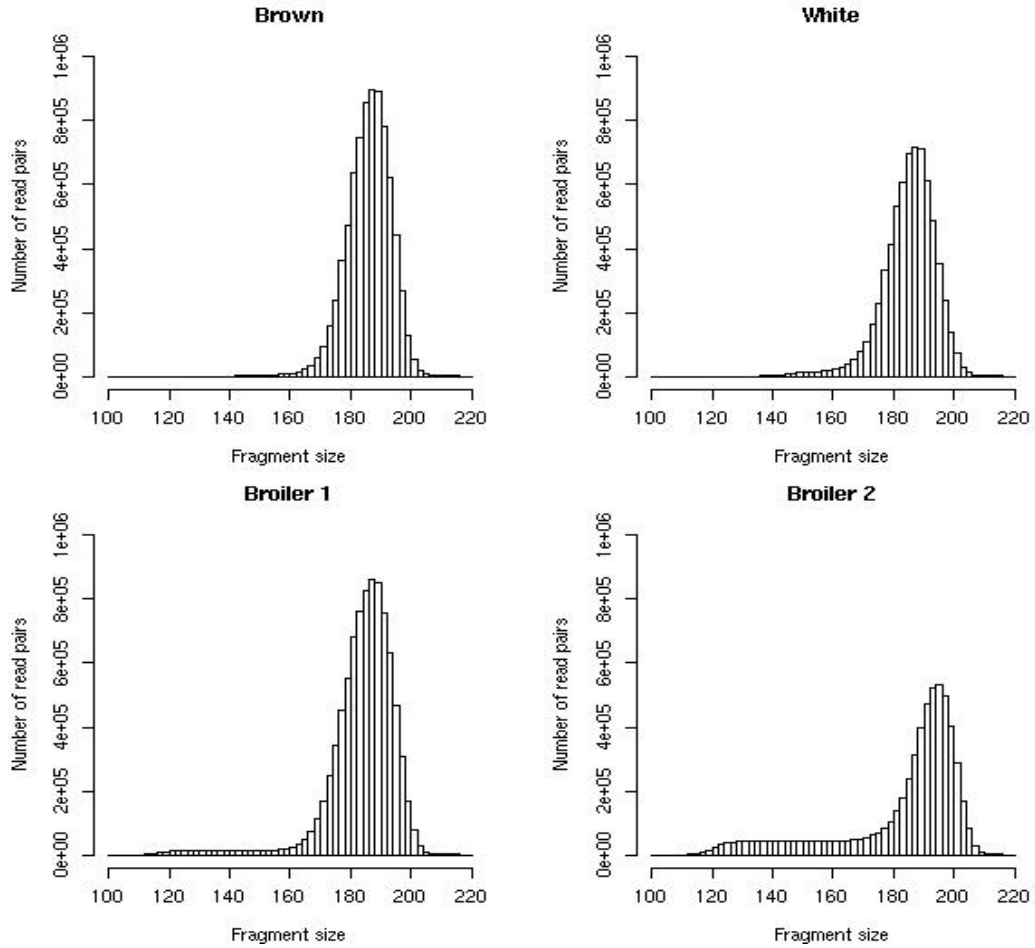

Figure 1: Distribution of fragment sizes for concordantly mapping reads in the four sequenced chicken breeds. For unclear reasons, broiler 2 had remarkably higher representation of smaller fragments (left long shoulder), whereas fragments in base pairs of the size range 180-200 were two magnitudes less abundant compared to the three other breeds.

Because of the high number, not all of the clusters are presumed to represent a true genomic rearrangement, but are incorrectly mapped reads caused by sequencing errors that result in low quality mapping. Therefore, the average mapping quality of discordantly mapping read pairs was evaluated per chromosome compared to the average mapping quality scores of read pairs that mapped consistently within the reference genome. However, the average mapping quality of discordantly mapping reads was similar to the mapping quality of concordantly mapping read pairs (Table 3). We also observed that the average coverage by paired reads differed up to two-fold between chromosomes, but the number of fragments per chromosome in the RLL highly correlated with chromosome size.

Table 3: Comparison of the mapping quality and distribution between concordantly and discordantly mapping read pairs.

Chromo -some	*Number of mapping read pairs*		*Average mapping quality*		*Mapping density*		*RRL density*
1	5329141	*15630*	67.92	*69.11*	38	*12860*	*1148*
2	3968343	*15049*	68.14	*71.29*	39	*10291*	*1149*
3	3344481	*11031*	68.87	*68.20*	34	*10303*	*1119*
4	2758645	*8155*	68.53	*70.40*	34	*11555*	*1098*
5	1975228	*5390*	68.53	*67.93*	32	*11547*	*1065*
6	1258393	*2782*	68.31	*69.69*	30	*13443*	*1056*
7	1336228	*4669*	68.78	*65.41*	29	*8221*	*1053*
8	1119526	*2866*	68.63	*72.82*	27	*10702*	*1067*
9	1016524	*3232*	68.16	*69.65*	25	*7907*	*1028*
10	761372	*2725*	68.20	*69.52*	30	*8278*	*1044*
11	677920	*1381*	68.56	*68.70*	32	*15879*	*1050*
12	864303	*3039*	68.33	*69.74*	24	*6758*	*989*
13	780565	*2107*	68.47	*66.72*	24	*8976*	*966*
14	740461	*3512*	67.86	*69.36*	21	*4504*	*929*
15	669260	*1378*	68.56	*68.47*	19	*9411*	*916*
20	722054	*2501*	68.78	*68.27*	19	*5592*	*911*
Z	1845751	*11981*	68.05	*68.79*	40	*6227*	*1271*

The number of concordant and discordant (in italics) mapping read pairs per chromosome are given. The average mapping quality of concordantly and discordantly mapping read pairs was calculated per chromosome. By calculating the mapping density, the distribution of mapping read pairs over the genome were evaluated. Mapping density was calculated by dividing the chromosome length by the number of concordantly/discordantly mapping read pairs. RRL density was calculated to ascertain the contribution of the RRL

approach to differences in mapping density. RRL densities were calculated by dividing the chromosome length by the (in silico) estimated number of RRL fragments.

To be considered as a true putative SV cluster, we required both ends to have an average mapping quality similar to concordantly mapping reads, which was ~60. In total, 7,789 clusters consisting of 3794 deletions, 3931 insertions, and 64 inversion breakpoints met this criterion. SV clusters predicting a deletion or insertion were further prioritized for confirmation screening on the basis of parameters listed in the Methods section. To validate our approach for identifying SVs, we initially evaluated 15 (SV13-28) predicted SVs (Table 4) using PCR to genotype pooled samples from the four chicken breeds with primers spanning predicted breakpoint junctions. A total of eight SVs yielded a clear PCR product of the expected size (Figure 2A). For these SVs, PCR was performed on individuals from breeds in which the SV was confirmed to be present by the SV-specific PCR product (Figure 2B). Individual SV-specific PCR products typed homozygous for the SV were sequenced to disentangle the rearrangement at the base-pair level. The sequence analysis results for these eight identified rearrangements were all consistent with our SV predictions.

Discriminating putative SVs from false positives

The results suggest that the presence of concordantly mapping reads partly overlapping the predicted SV region did not correlate with the quality of SV prediction, whereas reference errors in the predicted SV region correlated negatively. Furthermore, the results indicate that putative SVs predicted by single or few discordantly mapping read pairs that mapped a slightly different distance than expected were false positives, whereas the majority of putative SVs with greatly deviating mapping distances were confirmed as being true SVs. With this limited number of observations, we formulated a simple but fitting rule to determine SV clusters with a high likeliness to represent a genomic rearrangement from false positives.

We hypothesize that the size range of targeted DNA fragments isolated from the gel might contain a very small fraction of fragments outside the established size thresholds (Figure 1). This lack of proper separation is likely caused by migration artefacts caused by secondary DNA structures. To compensate for this bias, we required that predicted SVs based on discordantly mapping read pairs that mapped to the reference between 220 and 720 bp meet a

representation constraint. In our proposed validation rule, we assumed an inverse relationship between the span-size deviation of a predicted SV and the number of discordantly mapping read pairs (n) required to predict a true SV. We hypothetically state that SVs meeting the abundance constraint (span-size deviation)×n >500 can be validated as true deletions.

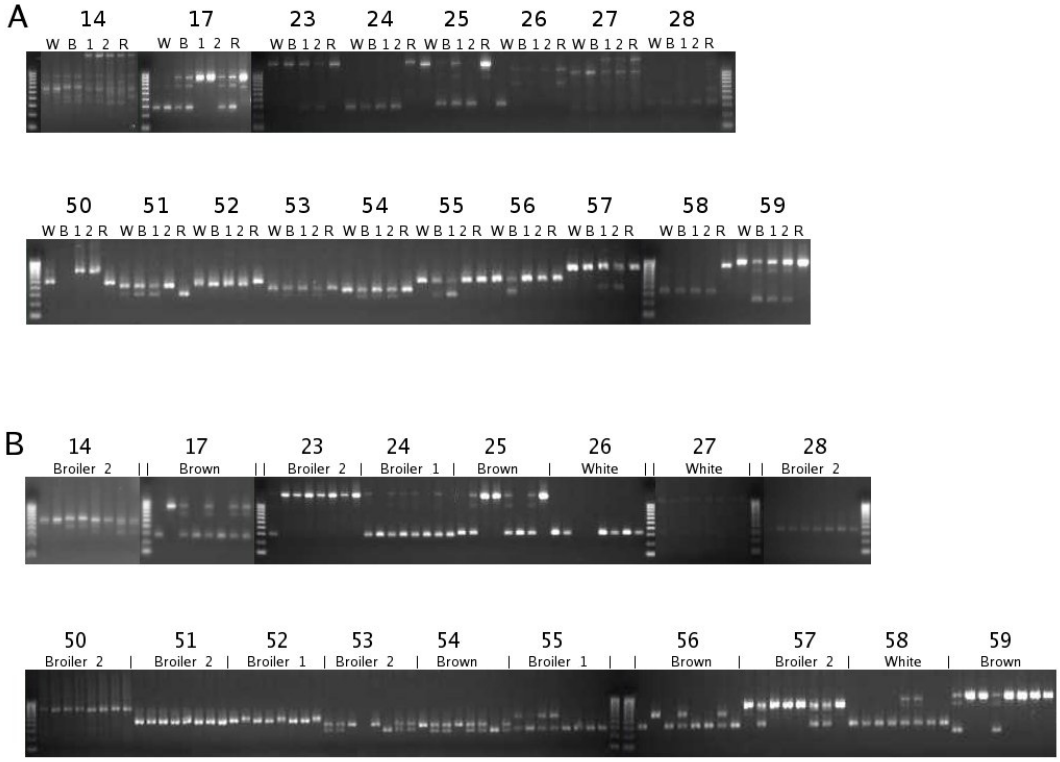

Figure 2: PCR-based genotyping on a breed level (A) and individual level (B). A) Genotyping for the presence of SVs in breeds, represented by pooled samples. Except for SV50 and SV51, a small (see Table 4 for approximate sizes and breed encoding) PCR fragment that was absent in the reference was expected in some of the breeds that have the deletion. In SV50 and SV51, a slightly larger PCR fragment than that observed in the reference was expected in breeds that have the insertion.
B) Genotyping for the presence of SVs in eight individuals of breeds in which the SV was detected in pooled samples. Except for SV50 and SV51, a small PCR fragment was expected in individuals homozygous for the deletion and SVs in which the reference genotype is too long for PCR. Heterozygous individuals in which both genotypes can be spanned (see Table 4) by PCR show two bands. In SV50 and SV51, both PCR fragments, which differ slightly in size, are expected in heterozygous individuals, whereas only the larger fragment is expected in individuals homozygous for the insertion.

Table 4: Validation structural polymorphisms.

	Prediction						*Confirmation*			
SV	***Span size***	***n***	***CMP***	***RE***	***aamq***	***Breed***	***Breakpoints***	***Size***	***Size in RRL***	***Breed***
15	251	1	X		97	2				
14	402	3			97	1,2	10_1627991-1628223	232	170	1,2
13	414	2			93	W				
18	640	1		X	99	1				
22	661	121	X	X	77	W,B,1,2				
17	729	4	X		94	W,2	3_110574268-110574832	564	165	W,2
20	780	6		X	96	W,1,2				
21	884	1	X	X	99	1				
19	970	2		X	99	1				
25	1248	3			73	2	1_188914114-188915200	1086	162	B,1,2
23	1319	1			97	2	2_55356006-55357163	1157	162	1,2
24	1376	2			70	2	4_23256240-23257477	1237	139	W,B,1,2
26	5845	1	X		90	W	2_112569238-112574924	5686	159	W
27	19574	15	X		96	W,1	-	-	-	-
28	8128	489	X		93	2	1_61836457_61844398	7941	187	W,B,1,2
50	64	48			71	B,1,2	2_152470660*			1,2
51	86	39			69	2	3_19576932	115	201	W,B,1,2
52	229	141			79	B,1,2	4_43663736-43663781	45	184	W,B,1,2
53	274	10			76	B,1,2	6_6687386-6687469	83	191	B,1,2
54	283	140	X		74	B,1,2	2_46860428-46860509	81	202	B,1,2
55	360	4			76	1	3_67474749-67474961	212	148	1
56	367	21	X		72	B	1_189692870-189693048	178	189	B
57	544	4			69	1,2	7_28561048-28561407	359	185	12
58	662	2			60	1	1_44948882-44949390	508	154	W,B,1,2
59	868	2	X		97	2	1_99177206-99177957	751	117	B,1,2

*Span size is the distance (in base pairs) on the reference sequence spanned by discordantly mapping read pairs. The number of observed discordantly mapping read pairs that support the presence of this structural variant (SV) is given by n, whereas CMP represents the number of concordantly mapping read pairs present in that particular genomic region. Discordantly mapping read pairs spanning an assembly problem in the reference genome are flagged in the RE column. The alternative mapping quality of a predicted SV is the average mapping quality calculated over discordantly mapping read pairs within a cluster. Deletion breakpoints are in the notation chr_start-stop, whereas insertion breakpoints are given in the notation chr_position. Breakpoints were not acquired for SV27 and not accurately acquired due to low sequence complexity in SV50. W = white egg layer; B = brown egg layer; 1 = broiler 1; 2 = broiler 2.*Due to the low sequence complexity, the exact location of insertion could not be revealed*

We assumed that this empirical rule is also applicable to insertions predicted by read pairs that map (too short) a distance of 32-100 base pairs. To test our empirical rule, we applied it to the subset of deletion (n=3794) and insertion (n=3931) clusters used in the previous validation study, obtaining 186 candidate putative deletions and two insertions. Both insertion candidates (SV50 and SV51) and a total of eight deletions (SV52-SV59), four of which narrowly met the rule constraints (Figure 3), were selected for confirmation. PCR-based genotyping analysis showed that all selected candidates were confirmed in the pooled samples (Figure 2A). We also observed that the PCR-based SV genotyping results for pools correlated well with the predicted presence of a particular SV in the breeds based on the sequence dataset (Table 4).

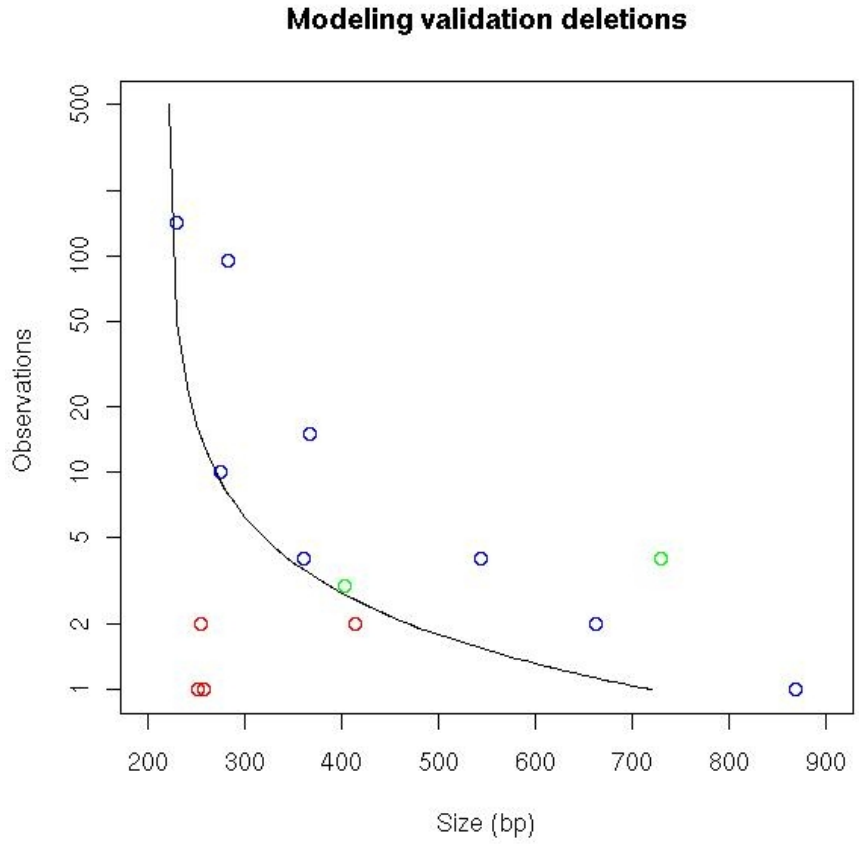

Figure 3: Distinguishing putative deletions from false positives in genotyping validation results obtained by PCR.Predicted deletions in the initial validation study that were confirmed are in green; those that could not be confirmed are in red. The black line represents the discrimination rule (span-size difference)×n >500, which is valid for 220-720 bp. The SV predictions that were selected based on the model and confirmed are in blue.

Breed-specific and shared SVs

Genotyping results suggested that the presence or absence of SVs in a particular breed is fairly well predicted by the sequencing data. Therefore, we further analysed 186 rearrangements (deletions) validated by our rule for breed specificity. We also analysed breed specificity for 280 putative deletions that resulted from applying a less stringent read mapping quality constraint, which was also applied in previous SV detection studies [19,25]. The results were compared by plotting both data subsets in weighted Venn diagrams (Figure 4). In the validated dataset of 186 deletions, we detected the most SVs in broilers, 114 in broiler 1 and 109 in broiler 2, whereas fewer SVs were detected in the layer lines, 60 in white egg layers and 85 in brown egg layers. Ten percent of the rearrangements were present in all four breeds. SVs detected in white egg layers were 23% breed-specific, and the other 77% were evenly shared with the other breeds. The brown egg layers had the fewest breed-specific SVs (18%) and shared a remarkably high percentage (65%) with broiler 1. Broiler 1 and broiler 2 showed similar percentages of breed-specific SVs, and 36% of the SVs in broiler 2 were shared with broiler 1. Applying a less stringent mapping quality constraint resulted in a 50% increase in SVs, whereas the distribution of SVs over the four chicken breeds remained approximately the same.

Distribution of predicted SVs

The majority of detected SVs were small (Figure 5); roughly 85% of all SVs were < 1 kb and 60% were < 500 bp. However, we also predicted and validated SVs spanning multiple kilobases. Predicted SVs validated by our rule were mapped to the chicken chromosomes, and we observed an even distribution (Figure 6). Sequence annotations of the regions overlapping the identified SVs were extracted from Ensembl [26]; 44% of the SV read pairs mapped within genes. The read pairs for a minor fraction of the SVs (~2%) spanned predicted exons; these were SVs putatively affecting genes at the transcriptional level (Table 5). The majority of all predicted SVs represented a putative deletion of low complexity and repetitive sequence motifs in intronic or intergenic regions (Table 6). A remarkable predicted SV was SV52, representing a deletion within exon ENSGALE00000116074 of gene ENSGALG00000010719, which has been annotated as DNA glycosylase FPG2.

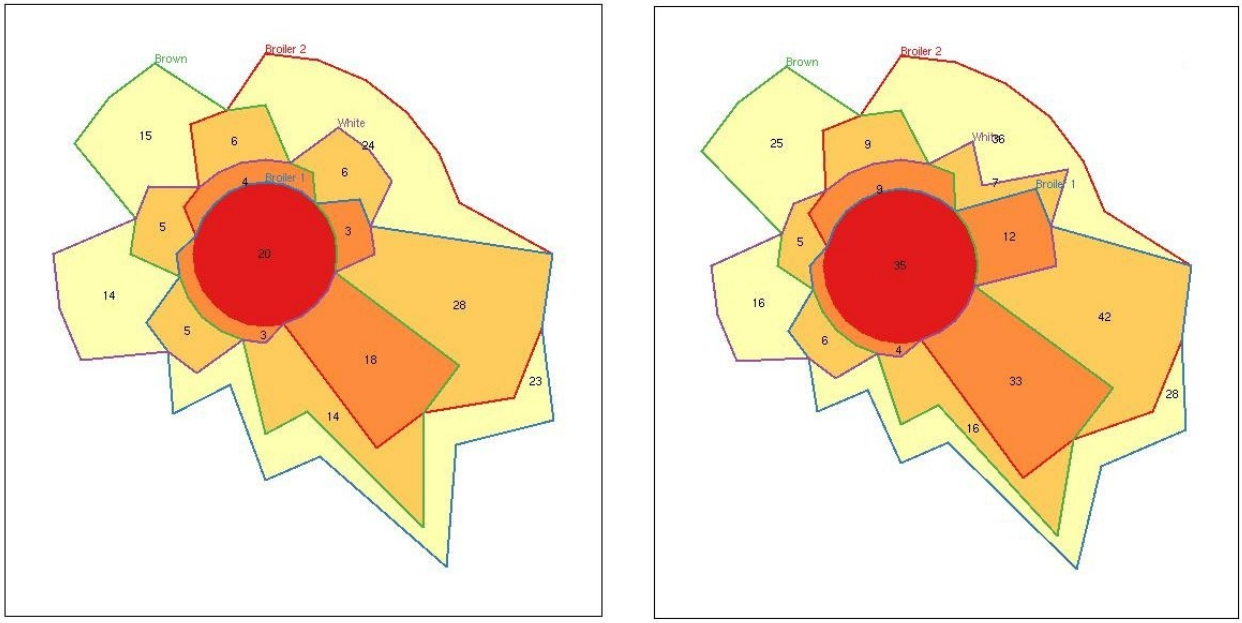

Figure 4: Venn diagrams representing the distribution of predicted deletions in the four chicken breeds at mapping constraints 60 (left) and 35 (right). The number of structural variants is proportionally represented per breed, and line colors were assigned as follows: green = brown egg layer; blue = broiler 1; red = broiler 2; and purple = white egg layer. For example, the area that is surrounded by the blue line in the left diagram represents SVs found in broiler 1. Of these, 23 were specific for broiler 1 (yellow area), and 28 were shared with broiler 2 (dark yellow area surrounded by both the blue and red lines). The orange area surrounded by the blue, red, and green line represent 18 SVs shared by broiler 1, broiler 2, and brown egg layers. The red area in the middle of the diagram surrounded by the four line colors represents 20 SVs shared by the four breeds analyzed.

SVs at base pair resolution and overlap with functional elements

All PCR-validated SVs were characterized by traditional sequence analysis to reveal their breakpoint locations, from which the chromosomal position and exact deletion/insertion sizes were derived (Table 4). Sequence losses were annotated using Ensembl [26]. For rearrangements in SV52, we analyzed the effect on the *in silico* transcript to which it was mapped. The majority of intronic deletions resulted in a loss of a variety of known repetitive motifs (Table 7). In contrast, we could not find annotations in Ensembl [26] for most losses in intergenic regions or known repeats using RepeatMasker (Smith and Green unpublished). DNA sequences at the SV breakpoints were analyzed for

signatures indicating the mechanism by which the SVs formed. We identified microhomology in three sequenced SVs. Finally, the SV we observed in a coding region involved a deletion in the end of the last exon (ENSGALE00000116074) of transcript ENSGALT00000038211.

Table 5: Annotation of putative deletions on the *in silico* transcript level.

Transcript(s)	*Modification*	*Protein*
ENSGALT00000005255	Truncation last exon	Flavin_mOase
ENSGALT00000003325	Truncation exon 9 or 5' deletion exon 10	PDZ domain
ENSGALT00000025445	5' deletion in last exon	Ionic channel
ENSGALT00000008864/40988	5' deletion in exon 4	Transcription factor
ENSGALT00000022933	Truncation exon 10	EGF-like
ENSGALT00000013428	Truncation exon 2	Unknown
ENSGALT000000002076/23151	Truncation last exon	ADP-ribosylation factor-like

Putative SVs with breakpoints predicted in exons were further analyzed in Ensembl [26]. Involved transcripts and protein functions were identified and putative modifications recorded.

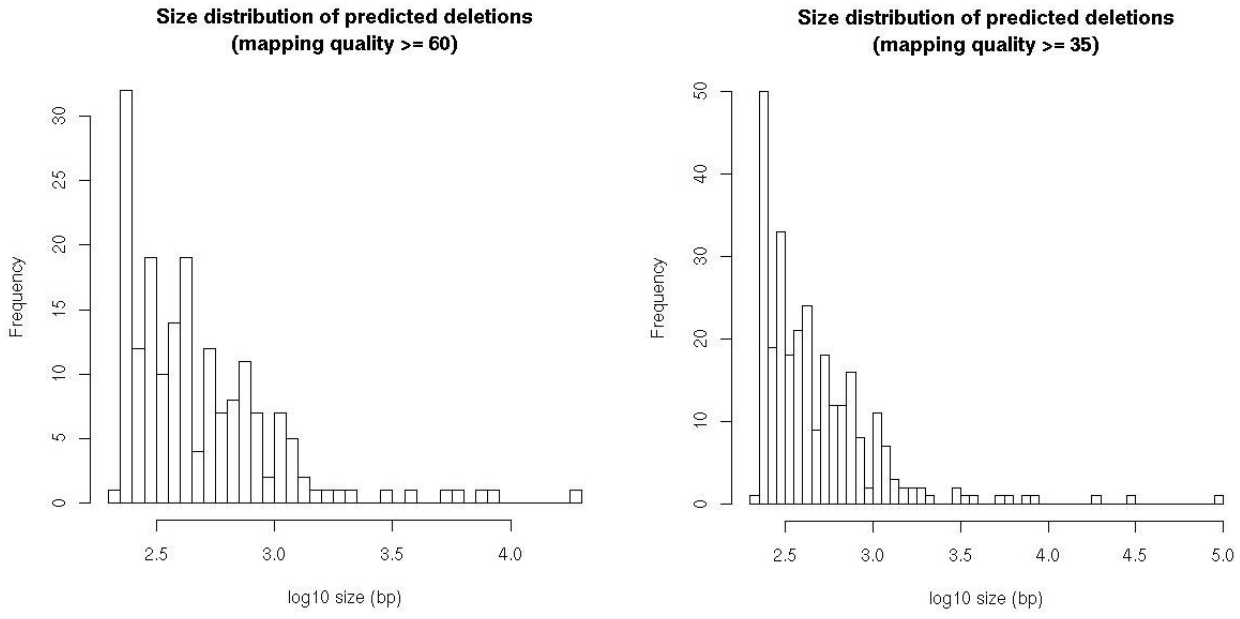

Figure 5: Size distribution of predicted deletions at two mapping constraints.

Discussion

By sampling a portion of the genome from four chicken lines using stringent SV detection constraints, we detected 188 SVs encompassing ~130 kb. Assuming considerable limitation in the detection of classes of SVs by our method, the chicken genome may differ in SVs to a greater extent than in SNPs. Therefore, we counted the total number of nucleotides involved. The majority of SVs identified by our method were small deletions, most of which resulted in a loss of repetitive motifs in intronic regions or a loss of unannotated sequences in intergenic regions. Both insertions mapped to intergenic regions as sequences of a few tens of base pairs and low complexity. However, we also predicted rearrangements in coding regions, revealed the exact breakpoints on the reference genome for 16 SVs, and confirmed our predictions. To what extent SVs in intronic and intergenic regions contribute to the evolution of the chicken genome or chicken phenotypes remains unclear, especially because the functions of these genomic regions are largely unknown [27]. To date, studies involving the detection and exploitation of genetic variation in chicken encompass large SVs by means of CNVs but do not include smaller SVs. Our study reveals that, given their high frequency, these smaller SVs will need to be incorporated in genotyping because they might explain phenotypic differences. In addition, our data suggest that structural variation has contributed to genetic differentiation among current domesticated chicken breeds and the Red Jungle Fowl, and might have played a role in chicken genome evolution.

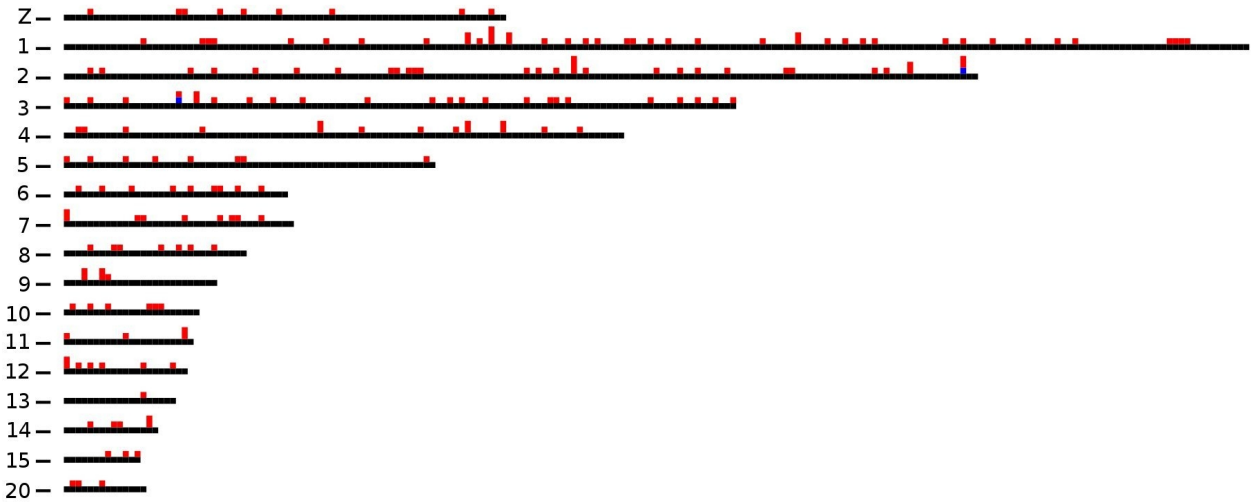

Figure 6: Distribution of predicted SVs over the chicken chromosomes.Shown are chicken chromosomes in which 186 deletions (red) and 2 insertions (blue) were identified.

Table 6: Putative functional annotations of predicted SVs.

		Coding			*Repeats*					
aamq	***n***	***% genes***	***% within exons***	***% exons***	***% CR1[1]***	***% GGLTR[2]***	***% other[3]***	***% TR[4]***	***% dust[5]***	***%![6]***
35	280	43.9	0.36	5	19.6	5.3	5.0	25.0	36.1	42.9
60	186	43.0	0.54	3.8	18.8	4.3	3.2	26.9	36.6	41.9

SVs of data subsets aamq 35 and aamq 60 were annotated based on their mapping location on the chicken genome. SVs were analyzed to determine whether they mapped within genes, within exons, or partially overlapped exons. 1CR1 = chicken repeat 1 [36]2GGLTR = Gallus gallus long terminal repeat3other = other specific repeat classes4SVs that mapped in repetitive sequences were analyzed for signatures of common repeats in the chicken genome and scanned for tandem repeats identified by Tandem Repeat Finder [37];5SVs that mapped in repetitive sequences were analyzed for signatures of simple repeats identified by the DUST algorithm [38]; 6The fraction of SVs that mapped in intronic and intergenic regions not identified as repetitive or low complexity are given in column "%!".

RRL-based approach to SV detection

Currently, sequence-based genome-wide surveys of SVs involve the preparation of whole genome fragment libraries in combination with paired-end sequencing. Such approaches require relatively large investments, particularly if multiple individuals from multiple breeds have to be screened. This study demonstrated the potential of massive parallel paired-end sequencing of RRLs constructed from the pooled DNA of multiple individuals. SVs were predicted based on the read pair information from the paired-end sequenced small insert RRL, which was purposely created for SNP detection. The small RRL size allowed for PCR-based confirmation and characterization of the SV at the base pair level of acquired deletions and small insertions with minimal sequencing efforts. Revealing inversion and translocation breakpoints is much more laborious due to the limited information RRL approaches provide. We showed that read pair analysis of a paired-end sequenced RRL is already sufficient for obtaining a first glimpse of SVs in a particular species. However, PEM of a randomly sheared and size-selected whole genome library provides a more complete catalog of rearrangements characterized between a sample and a

reference [1,19]. An even more complete picture including SVs of a larger size and more complex rearrangements will require paired-end sequencing of several libraries of different insert sizes [28]. Extremely demanding is the detection of all structural variation, which requires whole genome sequencing and *de no*vo assembly because the identification of (small) deletions and insertions with comparable or shorter length than the standard deviation of paired-end insert sizes cannot be identified by mapping approaches. Moreover, reference-based approaches, included mapping approaches, are biased to the completeness of the reference and, thus, ignore variants in regions that are missing from the reference genome due to structural variation. Finally, *de novo* assembly has the advantage of resolving SVs to a single base pair level, and inserted sequences can be obtained [29].

Next generation sequencing

We used a NGS approach to identify genomic rearrangements within four commercial chicken breeds by comparing their genomes to the sequenced chicken genome (Red Jungle Fowl). We excluded several classes of sequence reads from further analysis, including reads that did not show the restriction enzyme tag and those that showed more than one mismatch in the alignment. The first constraint was applied to eliminate false positive insertion predictions due to a breakdown of the RRL resulting in shorter spans of paired-end reads, whereas the second constraint was applied to reduce the number of false predictions due to sequencing errors. However, we realize that by taking these measures we also discard many read pairs because of true nucleotide variation, which occur in one of every 200 bp in the chicken [30]. The inclusion of read pairs with more than one mismatch in the alignment can be considered but has a risk of falsely predicted SVs due to mapping errors, requiring a revalidation of our proposed SV size deviation versus the observed frequency rule (Figure 3). On the other hand, reducing the mapping constraints might reveal additional true SVs potentially hidden in the considerable fraction of read pairs with only one end or no end mapped to the reference when using our mapping constraints. However, this fraction of read pairs with mapping problems might also largely represent sequences of gaps in the genome (estimated to encompass ~100 Mb in total) and, thus, cannot be mapped.

Table 7. Annotation of confirmed deletions and DNA signatures at breakpoints.

Nr	*Gene*	*Exons*	*Repeats*	*Signatures*
52	ENSGALG00000010719	ENSGALE00000116074		MH
54	ENSGALG00000012116			
53				
56				
55				
14	ENSGALG00000001729		trf1	MH
57	ENSGALG00000011699		dust	
58			dust	
17	ENSGALG00000016679		CR1-F0, Z-REP, trf, dust	
59				
25			dust, trf	
23	ENSGALG00000012402		dust, trf	
24	ENSGALG00000020249		dust, trf	
26			CR1-Y4, dust, trf	
28	ENSGALG00000012956		CR1-D2, Mariner1, GG, dust	MH

Deletions were annotated based on their mapping position on the chicken genome and deleted sequences were analyzed for common and more chicken-specific repeats. trf = repeats identified by Tandem Repeat Finder [37]; dust = simple repeats identified by the DUST algorithm [38]; CR1, = chicken repeat 1 [36]; Z-REP = macrosatellite family on chicken chromosome Z [39]; GG = repeats on the chicken genome identified by RECON [40]. We also analyzed the DNA sequence at SV breakpoints for signatures indicating the mechanism by which the SVs are formed, and we identified microhomology (MH) in some cases.

SV distribution across breeds

Theoretically, our approach for identifying SVs allows the prediction of SVs and insight into how a predicted SV is distributed across breeds. We showed that the observed distribution of SVs is a good predictor for the actual distribution of the SV in breeds. Even with limited sampling, predicted SV distributions correlated with the PCR-based genotyping results of pooled samples (Table 3). In general, PCR-based genotyping revealed that predicted SVs are more widely shared in breeds than predicted by our sequencing-based estimation. This situation is caused by limited sampling, and the reduction of target sequence complexity by creating RRLs might have contributed to this difference. Our sampling regimen required enzyme recognition sequences flanking a SV within the size range for the RRL to include a particular SV in the RRL. Breed-specific SNPs in AluI sites may have caused one or both SV alleles to not be sampled and are, thus, not predicted to be present in that breed, consequently affecting our sequencing-based estimation of SV distribution across breeds. Conversely, our PCR-based genotyping approach with pooled samples was not affected by sampling limitation or AluI SNPs and revealed the presence of SVs in a breed even at allele frequencies of 0.1 (data not shown).
Because of the difference in the predicted presence of a SV in a breed and the genotyping results, we realize that the 186 SVs with which we estimated breed specificity might not be fully representative. The use of different RRL sizes (150-200 bp in layers and 125-200 in broilers) is reflected in a 1.5-2-fold difference in the SVs detected in broilers and layers. The fairly large percentage of SVs shared in broilers can be interpreted as being due to the effects of selection during line development by commercial companies and is consistent with the results of recent SNP genotyping [31], but it might be over-estimated in our study due to the difference in RRL construction. The percentage of predicted SVs shared by brown egg layers and broiler 1, however, is an indication that these breeds are more genetically related compared to the other breeds. Recent SNP genotyping results for brown and white egg layers and three broiler lines also indicated that the brown egg layer breed is more closely related to broiler lines than to white egg layers [31], which is in agreement with our conclusion based on SV distribution.

Abundance, location, and size of SVs in the chicken genome

The reduction in the percentage of the genome covered by sequencing a RRL instead of randomly sampling the whole genome placed high constraints on the detection of SVs. The actual amount of SVs is likely much higher because we only sampled those that are flanked by restriction sites, and such that the intermediate sequence length of the variant was in the size range of the RRL. Large insertions were not expected to be detected because our RRL approach only allows for the detection of up to about 170 bp, the size between the maximum RRL fragment size (~200 bp) minus the mapping size of two completely overlapping reads (32 bp).

Although the larger SVs are most likely under-represented in our data due to the constraints of the applied detection method, we can conclude that the majority of SVs in the chicken genome are smaller than 1 kb (Figure 5). This finding is consistent with human studies [2] in which SV abundance inversely correlated with SV size. We observed that 99% of the predicted SVs were located in intronic (43%) and intergenic regions (56%), which together comprise ~90% of the chicken genome. As expected, SVs were less abundant in coding regions, which is consistent with the idea that the most common rearrangement mechanism requires substrates, such as microhomology, low copy repeats, and segmental duplications, which are more abundant in non-coding regions [10,32,33]. In 3 of 15 sequenced SV breakpoints, we were able to identify signatures in the DNA sequence indicating the mechanism by which SVs are formed. All identified signatures involved microhomology at the breakpoint junction that resulted from either nonhomologous end-joining or replication fork stalling and template switching events [34]. Other SVs did not show a clear sequence signature.

Conclusion

We provided a first glimpse of the abundance and genomic locations of structural variation in the chicken genome by identifying 188, mostly small, rearrangements, some of which were in coding regions, though a majority were located in non-coding regions. Based on the present data, we expect to find thousands of small (<1 kb) and hundreds of larger rearrangements in the whole chicken genome, encompassing more nucleotides than SNPs, and that are putatively involved in phenotypic variation. We observed that structural variation has contributed to genetic differentiation among current domesticated

chicken breeds and the Red Jungle Fowl. Finally, we showed that little sequencing effort on a reduced representation of a genome is sufficient for the detection and base pair level annotation of a variety of SVs in a sequenced genome.

Methods

SV detection using RRLs of pooled samples and NGS

Individual DNA samples were pooled according to breed and the genome complexity reduced by isolating a fraction of a complete genome digest. The isolated genome fraction was paired-end sequenced using Illumina genome Analyzer technology. The paired-end reads were aligned to the reference chicken genome and SVs identified as significant differences between the mapping distances identified by the paired-end reads and the size range used for constructing the RRLs. Deletions relative to the reference genome were identified by paired ends spanning a genomic region in the reference genome longer than the size in the RRL, whereas insertions were identified by paired ends spanning a shorter genomic region in the reference sequence than expected based on the RRL. Inversion breakpoints were detected by paired ends that mapped in a different relative orientation compared to the reference genome.

Paired-end sequencing

Genomic DNA was extracted from 300 µl of blood from 25 unrelated F_0 individuals from brown and white egg layer lines and two broiler lines consisting of 13 males and 12 females (Broiler 1) and 25 males (Broiler 2) using a Puregene DNA isolation kit (D-70KA; Gentra Systems, Inc., USA).
The RRLs were prepared by digesting 25 µg of pooled DNA using 1,000 units of the restriction enzyme *Alu*I in a total volume of 240 µl. The selection of the restriction enzyme was based on the 10-fold reduction of genome complexity in the optimum size range (100-200 bp) of the sequencing technology platform (Genome Analyzer, Illumina). The digested DNA sample was fractionated on polyacrylamide gel at 100 V for 3 h and stained with ethidium bromide. The size fractions were sliced out of the gel and the DNA recovered by shearing the gel pieces and eluting over night in 300 µl recovery buffer (8 mM Tris pH 8.0, 0.08 mM EDTA, 1.25 M ammonium acetate. After a 15-min incubation at 65°C, the eluent was purified using a Montage DNA Gel Extraction Device (Millipore Corporation, Bedford, MA) and precipitated with isopropanol. The DNA was

washed with ethanol and re-suspended in DNA hydration solution (Gentra Systems, Inc., USA).
We prepared the Genome Analyzer paired-end flow cell according to the manufacturer's protocol. Five picomole aliquots of the RRLs were processed using the Illumina Cluster Generation Station (Illumina, Inc., USA) following the manufacturer's recommendations. The Illumina GAII Genome Analyzer (Illumina, Inc., USA) was programmed to produce a theoretical fixed read length of 36 bp.
Images from the instrument were processed using the manufacturer's software to generate FASTQ sequence files. Paired reads that had both the RRL restriction tag and a per base phred (Ewing and Green, 1998) quality score of at least 20 were aligned to the chicken genome (WASHUC2) using the MAQ [35] algorithm v0.7.1.

Artefact removal

Paired reads in which one or both ends were mapped with more than one mismatch or mapped ambiguously on the reference sequence were excluded from analysis, as these would not reliably detect SVs. Discordantly mapping read pairs in which the two ends mapped >220 bp apart were classified as deletions and subsequently clustered based on overlapping mapping positions. SVs longer than 100 kb disrupted clustering and were excluded. Read pairs that mapped within 100 bp of each other were classified as insertions, whereas read pairs that mapped with one of the two ends in the incorrect orientation were classified as inversions. Both insertions and inversions were also clustered based on mapping positions.

Confirmation of identified SVs

For each SV cluster, we recorded the number of reads spanning the rearrangement, regardless of whether a normally mapping pair was observed or whether a sequence gap in the WASHUC2 build was present within the genomic range in which the deletion was predicted. SV clusters were prioritized for validation as follows: (i) an alternative mapping quality score of at least 60, (ii) both reads of a discordantly mapping pair mapped within a single predicted Ensembl exon or gene [26], and (iii) the genomic sequence flanking the SV allows primer design within 200 bp. We applied these criteria for selecting candidates distributed over the 220 bp-20 Kbp (deletions) and 32 bp-100 bp (insertions) size ranges. If these criteria yielded more than one candidate, the

candidate with the highest alternative mapping quality score was selected. Primers were designed to span the possible breakpoint by locating them 40-200 bp outside the mapping location of discordantly mapping read pairs. The minimum and maximum aberrant PCR product size was expected to be the sum of the minimum/maximum fragment size in the RLL and required flanking genomic region for primer development. PCR reactions were initially performed in the Red Jungle Fowl reference and the pooled samples of all four breeds. For breeds in which the rearrangements were detected, individual samples were genotyped by PCR. The PCR products of homozygous individuals, or samples in which only the aberrantly sized product resulted, were sequenced on a conventional Sanger capillary sequencer and the results compared to the reference sequence to identify breakpoints. Both ends of the PCR product on the reference (Red Jungle Fowl) were sequenced and mapped to the reference to ensure that it originated from the expected genomic position. Confirmed SVs were defined as those for which PCR reactions resulted in a distinct band in the expected size range in at least the breed for which the rearrangement was predicted and with no matching band in the reference (Red Jungle Fowl). The PCR results had to be supported by unambiguous sequencing data mapping confirming the rearrangement.

Availability and requirements

The SVs identified in this study that have not been confirmed and annotated at the base pair level are available upon request, awaiting a central repository of structural variation in genomes.

Authors' contributions

HHDK designed and developed the SV prediction method and wrote the manuscript. BWD and RPMAC prepared the samples and performed the initial validation and genotyping analysis. AV and RO selected the animals to be sequenced and collected the samples. MAMG and RPMAC coordinated and supervised experiment implementation and assisted in the preparation of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Mari Smits and Hendrik-Jan Megens for critically reading the manuscript and their helpful comments. This study was funded by European Union grant FOOD-CT-2004-506416 (Eadgene). Sequencing of the RRLs was funded by Cobb-Vantress Inc, USA and Hendrix Genetics, The Netherlands.

References

1.McKernan KJ, Peckham HE, Costa G, McLaughlin S, Tsung E, Fu Y, Clouser C, Dunkan C, Ichikawa J, Lee C, Zhang Z, Sheridan A, Fu H, Ranade S, Dimilanta E, Sokolsky T, Zhang L, Hendrickson C, Li B, Kotler L, Stuart J, Malek J, Manning J, Antipova A, Perez D, Moore M, Hayashibara K, Lyons M, Beaudoin R, Coleman B, Laptewicz M, Sanicandro A, Rhodes M, Vega FDL, Gottimukkala RK, Hyland F, Reese M, Yang S, Bafna V, Bashir A, Macbride A, Aklan C, Kidd JM, Eichler EE, Blanchard AP: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding.** *Genome Res* 2009, **19**:1527-1541.

2.Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318**:420-426.

3.Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.

4.Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, Ley TJ: **A high-resolution map of segmental DNA copy number variation in the mouse genome.** *PLoS Genet* 2007, **3**:e3.

5.Guryev V, Saar K, Adamovic T, Verheul M, Heesch SAACV, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, Hubner N, Cuppen E: **Distribution and functional impact of DNA copy number variation in the rat.** *Nat Genet* 2008, **40**:538-545.

6.Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, Dai F, Li Y, Cheng D, Li R, Cheng T, Jiang T, Becquet C, Xu X, Liu C, Zha X, Fan W, Lin Y, Shen Y, Jiang L, Jensen J, Hellmann I, Tang S, Zhao P, Xu H, Yu C, Zhang G, Li J, Cao J, Liu S, He N, Zhou Y, Liu H, Zhao J, Ye C, Du Z, Pan G, Zhao A, Shao H, Zeng W, Wu P, Li C, Pan M, Li J, Yin X, Li D, Wang J, Zheng H, Wang W, Zhang X, Li S, Yang H, Lu C, Nielsen R, Zhou Z, Wang J, Xiang Z, Wang J: **Complete**

resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* 2009, **326**:433-436.

7. Daines B, Wang H, Li Y, Han Y, Gibbs R, Chen R: **High-throughput Multiplex Sequencing to Discover Copy Number Variants in Drosophila.**. *Genetics* 2009, **182**:935-941.
8. Chen W, Swartz JD, Rush LJ, Alvarez CE: **Mapping DNA structural variation in dogs.** *Genome Res* 2009, 19:500-509.
9. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ: **Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease.** *Nat Genet* 2008, **40**:1107-1112.
10. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-454.
11. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK: **A high-resolution survey of deletion polymorphism in the human genome.** *Nat Genet* 2006, **38**:75-81.
12. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, Grassi AD, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315**:848-853.
13. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution.** *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.
14. Hollox EJ, Huffmeier U, Zeeuwen PLJM, Palla R, Lascorz J, Rodijk-Olthuis D, Kerkhof PCMVD, Traupe H, Jongh GD, Heijer MD, Reis A, Armour JAL, Schalkwijk J: **Psoriasis is associated with increased beta-defensin genomic copy number.** *Nat Genet* 2008, **40**:23-25.
15. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller H, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C,

Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Forti MD, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B, P GROU, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, Clair DS, Stefansson K: **Large recurrent microdeletions associated with schizophrenia.** *Nature* 2008, **455**:232-236

16. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays.** *Nat Genet* 2007, **39**:S16-S21.
17. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37**:727-732.
18. Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE: **A genome-wide survey of structural variation between human and chimpanzee.** *Genome Res* 2005, **15**:1344-1356.
19. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PAW, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40**:722-729.
20. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RPMA, Groenen MAM, Deryusheva S, Gaginskaya E, Carré W, Waddington D, Talbot R, Völker M, Masabanda JS, Burt DW: **Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution.** *BMC Genomics* 2008, **9**:168.
21. Skinner BM, Robertson LBW, Tempest HG, Langley EJ, Ioannou D, Fowler KE, Crooijmans RPMA, Hall AD, Griffin DK, Völker M: **Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis.** *BMC Genomics* 2009, **10**:357.
22. Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin C, Imsland F, Hallböök F, Andersson L: **Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens.** *PLoS Genet* 2009, **5**:e1000512.
23. Elferink MG, Vallée AAA, Jungerius AP, Crooijmans RPMA, Groenen MAM: **Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken.** *BMC Genomics* 2008, **9**:391.
24. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.

25. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6**:677-681.
26. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690-D697.
27. Mattick JS: **RNA regulation: a new genetics?**. *Nat Rev Genet* 2004, **5**:316-323.
28. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Cooley RN, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fajardo KVF, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang G, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST,

Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-59.

29. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing.**. *Genome Res* 2010*,* **20**:265-272.
30. International Chicken Polymorphism Map Consortium Wong GK, Liu B, Wang J, Zhang Y, Yang X, Zhang Z, Meng Q, Zhou J, Li D, Zhang J, Ni P, Li S, Ran L, Li H, Zhang J, Li R, Li S, Zheng H, Lin W, Li G, Wang X, Zhao W, Li J, Ye C, Dai M, Ruan J, Zhou Y, Li Y, He X, Zhang Y, Wang J, Huang X, Tong W, Chen J, Ye J, Chen C, Wei N, Li G, Dong L, Lan F, Sun Y, Zhang Z, Yang Z, Yu Y, Huang Y, He D, Xi Y, Wei D, Qi Q, Li W, Shi J, Wang M, Xie F, Wang J, Zhang X, Wang P, Zhao Y, Li N, Yang N, Dong W, Hu S, Zeng C, Zheng W, Hao B, Hillier LW, Yang S, Warren WC, Wilson RK, Brandström M, Ellegren H, Crooijmans RPMA, Poel JJVD, Bovenhuis H, Groenen MAM, Ovcharenko I, Gordon L, Stubbs L, Lucas S, Glavina T, Aerts A, Kaiser P, Rothwell L, Young JR, Rogers S, Walker BA, Hateren AV, Kaufman J, Bumstead N, Lamont SJ, Zhou H, Hocking PM, Morrice D, Koning DD, Law A, Bartley N, Burt DW, Hunt H, Cheng HH, Gunnarsson U, Wahlberg P, Andersson L, Kindlund E, Tammi MT, Andersson B, Webber C, Ponting CP, Overton IM, Boardman PE, Tang H, Hubbard SJ, Wilson SA, Yu J, Wang J, Yang H, Consortium ICPM: **A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms.** *Nature* 2004, **432**:717-722.
31. Megens H, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH, Hanotte O, Groenen MAM: **Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken.** *BMC Genet* 2009, **10**:86.
32. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80**:91-104.
33. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77**:78-88.

34.Lee JA, Carvalho CMB, Lupski JR: **A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.**. *Cell* 2007, **131**:1235-1247.
35.Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858.
36.Stumph WE, Kristo P, Tsai MJ, O'Malley BW: **A chicken middle-repetitive DNA sequence which shares homology with mammalian ubiquitous repeats.** *Nucleic Acids Res* 1981, **9**:5383-5397.
37.Benson1999 Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
38.Morgulis A, Gertz EM, Schäffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences.** *J Comput Biol* 2006, **13**:1028-1040.
39.Hori T, Suzuki Y, Solovei I, Saitoh Y, Hutchison N, Ikeda JE, Macgregor H, Mizuno S: **Characterization of DNA sequences constituting the terminal heterochromatin of the chicken Z chromosome.** *Chromosome Res* 1996, **4**:411-426.
40.Bao2002 Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.**. *Genome Res* 2002, **12**:1269-1276.

5 Genome wide SNP discovery and analysis in mallard (Anas platyrhynchos).

Hindrik H. D. Kerstens1*§, Robert H. S. Kraus2*, Pim Van Hooft2, Richard P. M. A. Crooijmans1, Jan J. Van Der Poel1, Johan Elmberg3, Alain Vignal4, Yinhua Huang5, Ning Li5, Herbert H. T. Prins2, Martien A. M. Groenen1.

1 Animal Breeding and Genomics Center, Wageningen University, Marijkeweg 40, Wageningen, 6709 PG, the Netherlands

2 Resource Ecology Group, Wageningen University, P.O. Box 47, 6700 AA, Wageningen, The Netherlands

3 Aquatic biology and chemistry, Kristianstad University, SE-291 88, Kristianstad, Sweden

4 UMR Génétique Cellulaire, Centre INRA de Toulouse, 31326 Castanet-Tolosan France

5 State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing 100094, China

*These authors contributed equally to this work

§Corresponding author

Email addresses:
HHDK: Hindrik.Kerstens@wur.nl
RHSK: Robert.Kraus@wur.nl
PvH: Pim.vanHooft@wur.nl
RPMAC: Richard.Crooijmans@wur.nl
JJvdP: Jan.vanderPoel@wur.nl
JE: Johan.Elmberg@hkr.se
AV: Alain.Vignal@toulouse.inra.fr
YH: yinhuahuang@126.com
NL: ninglcau@cau.edu.cn
HHTP: Herbert.Prins@wur.nl
MAMG: Martien.Groenen@wur.nl

Ready for Submission

Abstract

Background

Next generation sequencing technologies makes it feasible to economically obtain the genomic sequence information that is currently lacking for most economically and ecologically important species. For the mallard (*Anas platyrhynchos*), genomic information like sequencing data and single-nucleotide-polymorphisms (SNPs) is limited. The duck is, besides a species of agricultural importance, also a member of a taxon of focus when it comes to long distance dispersal of Avian Influenza. Therefore, it is important to facilitate duck breeding by providing sufficient genetic markers for QTL mapping and to study duck migration in a population genetic framework. For large scale identification of SNPs we performed sequencing of wild duck DNA and compared our data with ongoing genome sequencing of domesticated duck and data from a duck EST-sequencing project.

Results

More than a billion basepairs (bp) of high quality sequence information were generated resulting in a 16X coverage of a reduced representation library covering 5% of the wild duck genome (1.38Gbp). Sequence reads (62 bp) were aligned to a draft (domesticated) duck reference genome and allowed for the detection of over 122,000 SNPs within our (wild) duck sequence dataset. In addition almost 62,000 nucleotide positions on the domesticated duck reference showed a different nucleotide compared to the wild duck sequence reads on that position. Of SNPs that were identified within our sequence data of wild duck, more than 20,000 were shared with SNPs identified in the sequenced domestic duck or EST sequencing projects. The shared SNPs, being predicted in independent projects, were considered to be highly reliable and to represent true nucleotide polymorphisms and were used to benchmark non-shared SNPs on metrics like transition/transversion rates and the distribution of the SNPs over the nucleotide positions in the sequence reads. A comparison of the shared SNPs with the sequenced chicken genome indicated a uniform distribution of the SNPs across the duck genome. Genotyping of a representative sample of 364 shared SNPs resulted in a SNP conversion rate of 99.7%. The correlation of the minor allele count and observed minor allele frequency in the SNP discovery pool for the validated SNPs was 0.72.

Conclusion

We identified almost 150,000 SNPs in wild mallards that will likely yield good conversion rates in genotyping. Of these, ~101,000 SNPs were detected within wild duck and ~49,000 were detected as polymorphisms between wild and domesticated ducks. In the ~101,000 SNPs we found a fairly large subset of ~20.000 SNPs in common between the wild duck and the sequenced domesticated duck suggesting a low genetic divergence between wild duck and its domesticated relative. This subset showed almost 100% SNP conversion rate at genotyping. Comparison, on metrics like transition/transversion rates and SNP distribution over sequence reads, of the total SNP set (122,000 + 62,000 = 184,000 SNPs) to the validated subset, shows similar characteristics for both sets. This indicates that we have detected a large amount (~150.000) of accurately called duck SNPs, that will be beneficial for both industrial and ecological applications.

Background

The mallard (*Anas platyrhynchos*) is the most abundant and well-known waterfowl species. Besides duck is important as a hunted game species, it can also be seen as a flagship species in wetland conservation and restoration.Ducks (Anseriformes: Anatidae) in general have become the migratory taxon of focus when it comes to long distance dispersal of Avian Influenza in the wild [1,2], and the mallard has been identified as the most likely species that transmits avian influenza viruses [3,4]. The general migration trend of the mallard is from the north (breeding grounds) to the south (wintering grounds) avoiding freezing conditions at breeding sites [5], but in Europe no clear flyways could be inferred so far [6]. This is why Wink [7] proposed the use of SNPs to study bird migration in a population genetic framework. Since the number of SNPs necessary to detect low levels of differentiation is expected to be high (>80) for highly mobile organisms [8,9], we aimed at a high throughput discovery of SNPs in wild duck. Large scale discovery of SNPs in the genome of the wild mallard might also provide a useful set of markers in the descendant domesticated duck (*Anas platyrhynchos domestica*). With it being the third most consumed species of the world wide poultry market [10], the duck provides a valuable subject for detailed genomic studies. Nevertheless, genomic information about the duck is limited to a few studies providing only low resolution linkage and physical map [11,12]. Therefore our study of mallard will also facilitate duck breeding objectives by providing sufficient markers for improving the duck linkage map and allowing QTL mapping using SNPs. A

general limitation in developing a SNP-set in non-model organisms has been the unavailability of extensive sequence information from multiple individuals that represent a sufficient portion of the genetic variability of the population or species under study. However, the Illumina Solexa Genome Analyser technology [13,14,15] coupled with the approach of generating a reduced representation library (RRL) [16] was shown to be an efficient approach in solving this problem in turkey (*Meleagris gallopavo*) [17] and great tit (*Parus major*) [18]. Also in rainbow trout [19], pig [20,21] and cattle [22] next generation sequencing of RRLs has been effective in the identification of considerable numbers of SNPs. Here, we describe the discovery of more than 180,000 novel SNPs in the genome of the mallard, currently lacking a published sequenced genome. By lack of a reference genome we initially aimed for paired-end sequencing on a Illumina GAII of a library of fragments in the size range of 110-130 bps and with a read length of 76 bases. This creates an overlap between the forward and reverse reads of a pair which allows merging of the reads. Merging the reads helps in providing sufficient flanking sequence of a SNP which is a requirement for genotyping and is hard to retrieve in the absence of a reference genome. However, at the time of our study, sequencing of the duck genome and *de novo* assembly was in progress and almost completed by the Beijing Genome Institute (BGI). This allowed for SNP discovery by next generation sequencing of a RRL of pooled wild duck samples and mapping of almost 13 million of the resulting reads to a draft duck reference sequence. SNPs identified were compared with those observed within the reference sequence of domesticated duck (Huang et al., in prep.) and mallard EST sequencing (Vignal, unpublished data) resulting in more than 20,000 shared high quality SNPs.

Results

Complexity reduction

We targeted for a sequencing depth of about 40 times at limited sequencing cost by sequencing a fraction, or RRL, representing 5% of the duck genome. Restriction enzymes were screened for suitability for RRL construction, with the goal of a 20 fold complexity reduction of the duck genome within the targeted size range of 110-130 bp. Restriction analyses showed that these requirements are met by combining two libraries, one created by digestion with *Alu*I and one by digestion with *Hha*I, representing 4% and 1% of the duck genome, respectively.

An *in silico* digest of the highly similar chicken genome [23,24] predicts similar genome fractions of the RRLs of 4.1% for *Alu*I, but only 0.2% for *Hha*I (data not shown). After enzyme selection, we prepared two pooled DNA samples of nine wild mallard individuals from three locations across Europe. To prepare the libraries, we digested these samples with *Alu*I or *Hha*I and isolated fragments in the 110-130 bp size range from a preparative polyacrylamide gel. The genomic libraries were combined in the sequencing sample preparation procedure. By lack of a reference genome we aimed for paired-end sequencing on a Illumina GAII of the combined libraries and with a read length of 76 bases. This creates an overlap between the forward and reverse reads of a pair which allows merging of the reads. Merging the reads helps in providing sufficient flanking sequence of a SNP. This sequence is necessary for genotyping and is hard to retrieve in the absence of a reference genome. Merged paired reads, possibly supplemented with single reads, are subsequently clustered for SNP discovery.

Illumina sequencing and SNP detection

We generated 34.8 million 76 bp reads using three lanes on an Illumina GAII of which two lanes were run in paired-end mode. It was shown that a phred quality score [25] threshold of 12 ensures sufficient quality reads for SNP detection purposes [20,26]. Because the average base quality score over all reads dropped below 12 after position 62, reads were trimmed to 62 bp. After trimming, we performed additional quality based filtering (see methods) and finally we retained 16.6 million reads (47%) of 62 bp length corresponding to a total of 1.03 billion bp of sequence information (Table 1).

Table 1: Summary of DNA sequence filtering results

	raw (76 bp)	***l62 n. q12 o152[1]***	***%***	***paired-end***	***%***	***single***	***%***
reads	34,818,352	16,611,852	47.7	10,793,170	65.0	5,818,682	35.0
bases	2,547,361,732	1,029,934,824	40.4	669,176,540	65.0	360,758,284	35.0

Paired and single sequence reads remaining after filtering raw reads.1Raw sequences were filtered for length 62, without base-call errors (n or .). Singly represented reads are required to have a per base-call quality of 12. Sequences more than four times overrepresented, based on the raw RRL coverage (38X, see methods, were discarded.

Of these reads 35% were single and 65% were paired reads. By creating RRLs 5% (69 Mb) of the mallard genome was represented (estimated size 1.38 billion bp) based on several entries in the Eukaryotic genome size databases [27]. From this we calculated that the raw sequencing data is covering the sequence target 38 times whereas the quality filtered data provide a ~16x target coverage. Using MAQ [28] 12,823,563 of the reads could be mapped onto the duck reference genome (Huang et al., in prep.). A total of 632,163 putative SNPs were identified by MAQ [28] of which 122,413 candidate SNPs passed our applied SNP calling quality thresholds (see methods). This set of SNPs is further referred to as duck-RRL (d-RRL).

SNP usability

More than 98.8% of the SNPs were flanked by at least 40 bp on either side and do meet the requirements for probe design constraints for all genotyping platforms whereas all SNPs meet the flanking sequence requirements for an iSelect (Illumina) genotyping assay. For the 2565 SNPs that showed more than two alleles we only considered the most frequently observed minor allele because tri- or tetra-allelic SNPs are very rare and it is likely that other minor alleles represent sequencing errors instead of true variants. Analysis of the estimated allele counts of the SNPs in our dataset (Figures 1A and 1B) show that we obtained a majority of SNPs with a high minor allele count (MAC).

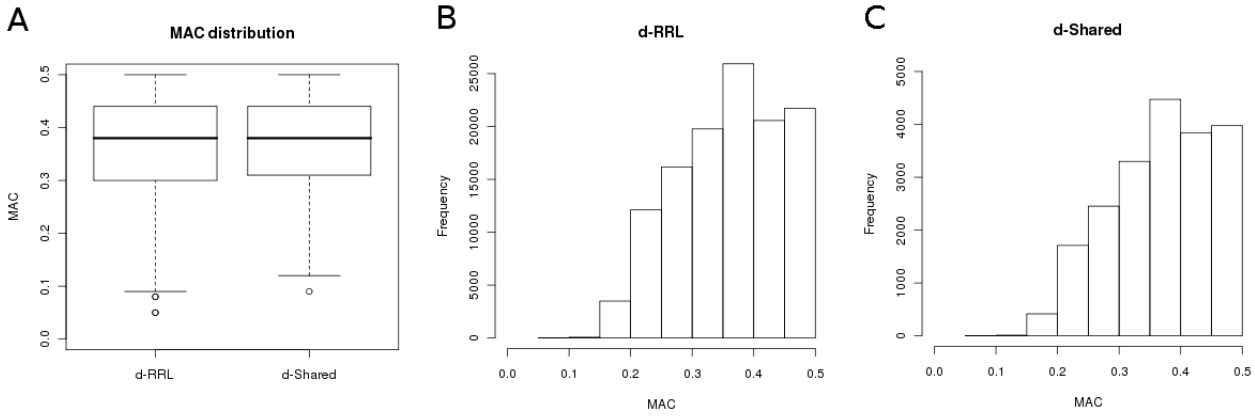

Figure 1:Minor allele frequency distributions.n the boxplot (A) MAC distributions of d-RRL (SNPs identified in this study) and d-Shared (SNPs that d-RRL shares with d-EST and d-WGS (also see Venn diagram Figure 2D)) are compared. Histograms (B and C) show MAC distributions of d-RRL and d-Shared at a bin width of 0.05

SNP quality assessment

Sequencing errors are more abundant in lower quality tailing nucleotides in next generation sequencing reads and putatively cause false SNP predictions. An increase in the number of SNPs towards the end of the reads is expected if sequencing errors are the cause of a substantial amount of predicted SNPs in the dataset. To validate our sequence filtering and SNP detection constraints, we plotted the distribution of the SNPs over the 62 positions in the sequence reads (Figure 2A). Positions one, two and 62 all show an underrepresentation of SNPs whereas positions three, four and five show an overrepresentation. SNPs are equally distributed over read positions 6 to 25 and at 26 the number of SNPs per nucleotide position drops but after this remains more or less steady till position 62. Because of the RRL insert size (~110-130bp), there is a putative overlap between paired forward and reverse reads (62 nucleotides each) from position 48 onwards . This putative overlap from bases 48-62 results in a higher sequence depth and a tiny increase in the number of SNPs being detected at these nucleotide positions (Figure 2A).

Sequencing errors result in the introduction of random polymorphisms resulting in an expected transition (A/G or C/T) versus transversion (A/C,A/T,C/G,G/T) rate of 1:2. However, in the d-RRL dataset we observed in the class transitions (TS) that the number of A/G substitutions almost equals the number of C/T substitutions. Also the substitutions within the class transversions (TS) occurred in comparable frequencies (Table 2). The TS:TV ratio for d-RRL is 2.3:1 which is similar to the ratio of 2.2:1 calculated for chicken, based on more than 3 million chicken SNPs present in dbSNP.

Sequencing errors were also evaluated per read position by plotting the TS:TV ratio observed over the 62 positions in the sequence reads (Figure 2). We observed steady expected TS:TV ratios for positions 7-61 whereas TS:TV ratios for positions one to six were lower and the TS:TV ratios for position 62 was higher than expected.

SNP benchmarking

At the time of our study, sequencing of the duck genome and *de novo* assembly was in progress and almost completed by the Beijing Genome Institute (BGI). Next generation sequencing data, covering both alleles of a single domesticated duck resulted in the identification of 2,826,871 putative SNPs (further referred to as d-WGS). Duck-EST sequencing identified a total of 6456 SNPs (further referred to as d-EST), in coding regions of the duck genome (A. Vignal, unpublished data).

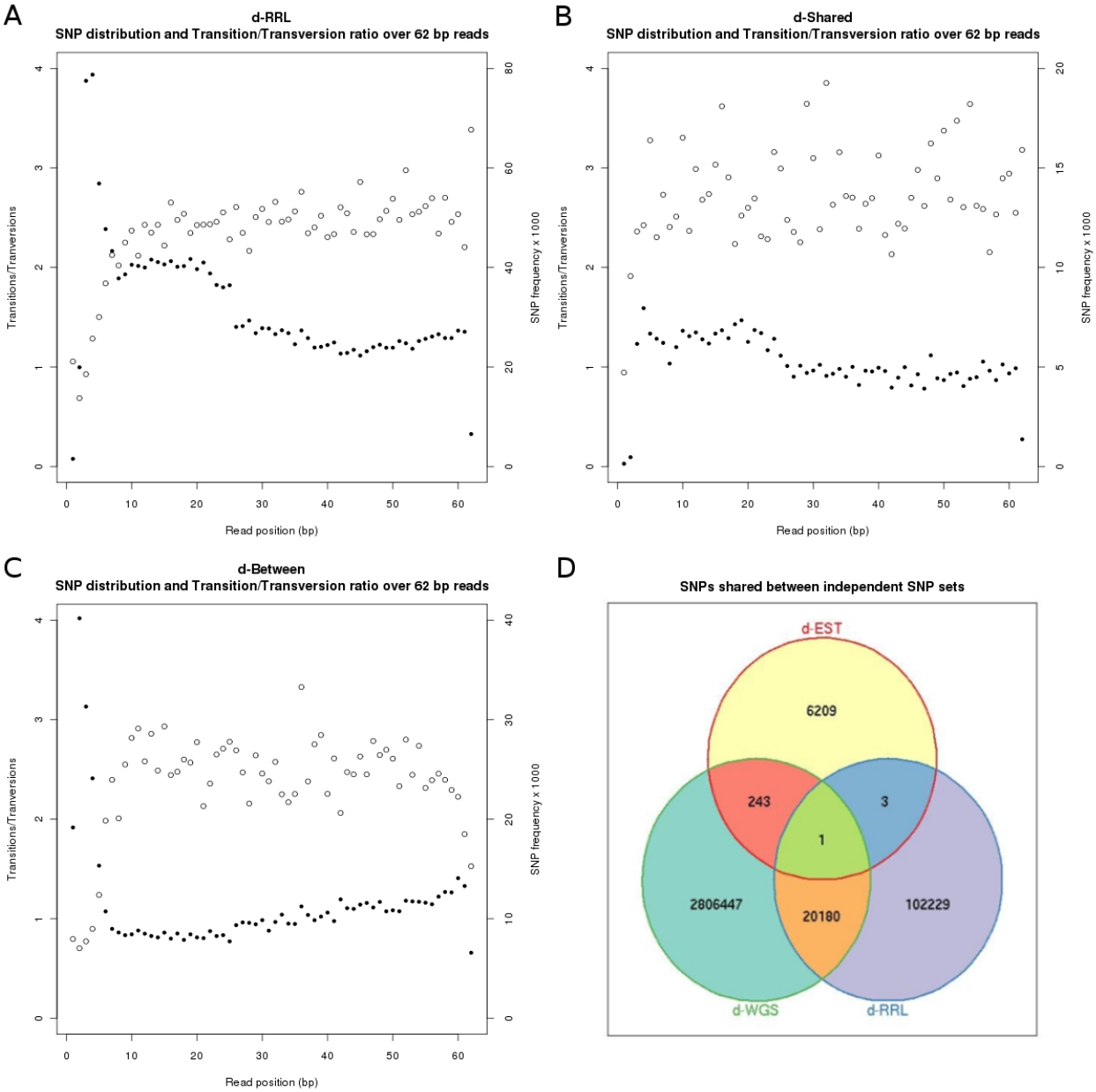

Figure 2: SNP distributions within datasets and between datasets. Diagrams A-C show the distribution of SNP predictions over the nucleotide position in the sequence reads for d-RRL, d-Shared and d-Between. Each filled dot represents the cumulative number of occurrences that the read position was involved in a SNP call. Open dots represent the average TS:TV ratio of SNPs indentified in that read position. Diagram D shows how many SNPs are shared between independent SNP sets d-EST (SNPs identified by EST sequencing of domesticated duck (Vignal, unpublished data)), d-WGS (SNPs identified in whole genome assembly of domesticated duck (Huang et al., in prep.)) and d-RRL (SNPs identified in RRL sequencing of wild duck (this study)).

To benchmark d-RRL we compared it with these two external and independent datasets and identified SNPs that are shared with either d-WGS or d-EST. We observed 20,180 SNPs (16.5%) in common between d-RRL and SNPs in the d-WGS dataset. Furthermore d-RRL had four SNPs in common with d-EST whereas d-WGS shared 244 SNPs with d-EST (Figure 2D). Only a single SNP was shared between all three datasets. The subset of SNPs (n=20,184) that d-RRL shared with either of the two other SNP resources is further referred to as d-Shared. We analysed d-Shared by calculating the MACs and the TS:TV ratios (Figure 1C and Table 2). Furthermore we plotted the TS:TV ratio per read position and the distribution of the SNPs over the 62 nucleotides of the sequence reads analogous as was done for d-RRL. In d-Shared we observed a similar distribution of MACs compared to d-RRL (Figure 1C). The distribution of the SNPs in d-Shared over read positions 7-62 is similar to that observed for d-RRL; however, d-Shared shows a higher variation in SNPs between the read positions (Figure 2B). Also TS:TV ratios at these read positions were similar with slightly more variation per read position in d-Shared.

Table 2 Transition/transversion ratios in SNP subsets

	Transitions		***Transversions***				***Total***	***TS:TV[1]***
subset	R	Y	M	W	S	K		
d-RRL	42313	42602	9658	9051	9114	9675	122,413	2.3
d-Shared	7300	7442	1396	1227	1334	1484	20,184	2.7
d-Between	20156	21333	5464	5165	4804	4830	61,752	2.0

[1]*=The transitions total divided by the transversions total for a data subset.*

Although reduced, also d-Shared showed a peak in SNPs on read positions three to six like we observed in d-RRL. However, TS:TV ratios for these positions were at expected level of >2.3 indicating that most SNPs in these read positions likely resulted from true nucleotide polymorphisms. Finally, compared to d-RRL, the d-Shared subset of SNPs showed a higher average TS:TV ratio of 2.7 and indicated a relative increase of (C/T) over (A/G) transitions (Table 2).

Domesticated versus Wild

Besides the identification of SNPs within the wild mallard population we also mined for nucleotide positions in the genome that show differences between the wild mallard population and the domesticated duck reference. We considered

nucleotides in the wild mallard consensus where MAQ [28] did not assign an ambiguity code but that were different from the corresponding non-polymorphic position in the domesticated duck reference. We identified 61,752 SNPs (further referred to as d-Between) and assessed the quality of this set of SNPs by plotting the TS/TV ratio per nucleotide position and plotting the distribution of the SNPs over the 62 nucleotide positions in the sequence reads (Figure 2C). The number of SNPs predicted in the first six read positions showed a high peak whereas from position six to 62 the number of SNPs per read position is more or less constant, steadily increasing towards the end. The TS:TV ratios were as expected except on the first six read positions and the tailing end, where it was lower than expected. Compared to d-RRL and d-Shared the overall TS/TV ratio of d-Between is lower, 2:1, and shows a relative increase of (C/T) over (A/G) transitions (Table 2).

The distribution of SNPs over the genome

Knowing genomic positions of SNPs as genetic markers is important. Many population genetic and genetic mapping applications rely on unlinked markers. Thus, for future use in generating a duck linkage map and performing QTL studies in duck it is essential that the SNPs are widely distributed over the genome. The next generation sequenced duck genome assembly that we used as a genome reference, consists of thousands of scaffolds and contigs which are not assigned to chromosomes. Estimating the distribution of SNPs across the duck genome therefore is not possible using this sequenced duck reference. For this reason the closest related available genome sequence (chicken) was used for estimating the distribution of the identified SNPs. Common and high quality duck SNPs (d-Shared) were aligned to the chicken genome and the distribution of this SNP-set was plotted over the chicken chromosomes (Figure 3). A total of 4,272 SNPs could be mapped to unique locations evenly distributed over the chicken genome.

SNP validation by genotyping

The d-shared subset of SNPs was validated by genotyping an animal panel consisting of 765 ducks using 384 predicted SNPs distributed uniformly over the chicken genome (Figure 3). A total of 364 (95%) SNPs gave reliable genotypes in the assay, and 363 (99.7%) of these were polymorphic. The average minor allele frequency (MAF) was 0.32 in the animals that made up the discovery panel and 0.31 in the whole animal panel (Figure 4). The average heterozygosity was 0.39 in the discovery panel and 0.34 in the whole animal

panel. The allele frequencies of polymorphic genotyped SNPs in the discovery pool showed a correlation of 0.72 with the allele counts derived from the sequence data in the nine animals represented.

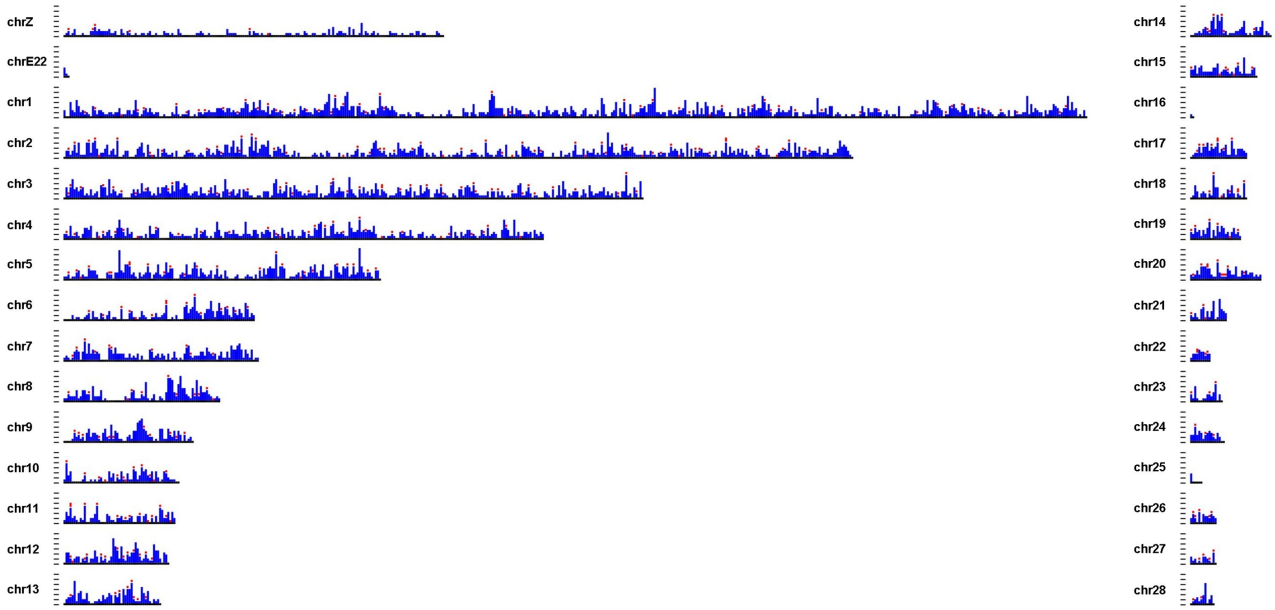

Figure 3:Distribution of duck SNPs that uniquely mapped on the chicken genome.In blue are duck SNPs that mapped uniquely to the chicken genome. Mapped SNPs that were selected for genotyping are in red. On the X-axis, the chicken genome in 400 kb intervals. On the Y-axis, the frequency (0-15) of mapped duck SNPs for a specific chicken genome interval is given.

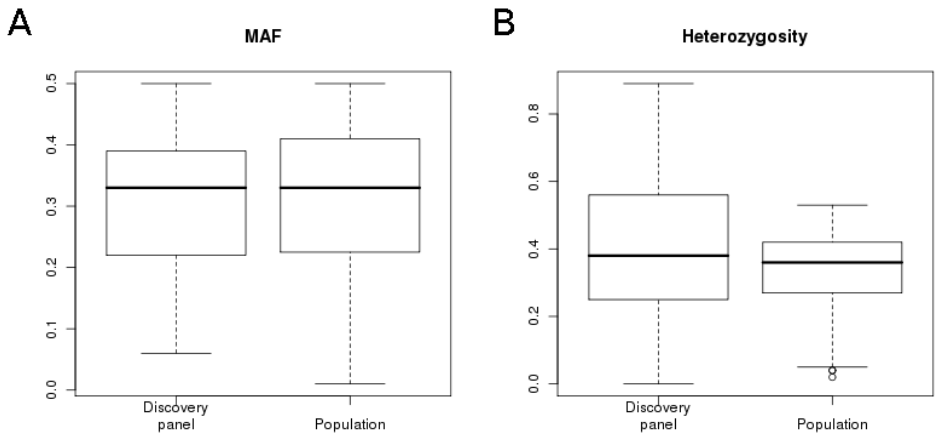

Figure 4: Genotyping minor allele frequency and heterozygosity distributions. Validation of the d-Shared subset involved genotyping of 384 selected SNPs on 765 ducks including the nine animals that made up the SNP discovery panel. Minor allele frequency (MAF) and heterozygosity of SNPs were calculated for the discovery panel as well as for the whole set of genotyped ducks.

Discussion

This SNP detection study is the first large sequence variant discovery performed in mallards. The availability of a large number of SNPs provides sufficient markers to study bird migration in a population genetic framework [7]. This large number of accurately called SNPs will also facilitate improved linkage maps of the duck genome and provide a sufficiently dense marker map allowing for high resolution QTL studies in duck, further facilitating duck breeding. Furthermore, such high density linkage maps are essential for chromosomal assignment of the sequence scaffolds of the sequenced reference genome.

SNP detection within a pool of wild European mallards

Initially, our study was designed to detect SNPs within a pool of wild European mallards by single-end and paired-end sequencing of a small fragment RRL. We targeted for genome libraries of sufficiently small fragments for paired reads to overlap. This allows the reads to be merged resulting in the complete sequence of the majority of the fragments in the RRL. Merged paired reads subsequently would serve as a reference genome. However, with the recent availability of a next generation sequenced duck genome assembly, a reference based mapping approach became feasible, enabling a more efficient SNP identification approach. This study shows that the overlap in the in general lower quality tailing end of paired-end sequence reads is beneficial in reference based SNP detection. We observed that the number of SNPs being predicted per read position shows a tiny increase in the overlapping tailing ends whereas earlier studies [17,18,20] reported decreasing numbers of predicted SNPs per nucleotide position towards the end of sequence reads. The TS/TV ratio of SNPs predicted in the overlapping tailing ends remains in the expected range (Figure 2A) suggesting that predicted SNPs are reflectig true nucleotide polymorphism. A local decrease in TS/TV ratio would be observed if SNPs in read positions (51-61 in d-RRL and 52-60 in d-Between) are caused by randomly introduced polymorphisms (e.g. sequencing errors). Because transitions are outnumbering transversions, coinciding with the idea that most SNPs are caused by CpG DNA methylation, we expect that the predicted SNPs represent true nucleotide polymorphisms. The increased number of SNPs at the overlapping tailing ends can be explained by a local higher sequence coverage, caused by sequence overlap of paired reads, resulting in a higher representation of variants. A higher coverage allows for multiple observations of the variant in low quality sequence allowing it to pass MAQ's quality thresholds to call it a

true SNP [28]. As a result, even more rare sequence variants in these overlaps will meet the minor allele occurrence constraint in the SNP detection method. An indication that the additionally identified SNPs at the read tailing ends involve rare sequence variants is the lower representation of these SNPs in d-Shared.

Ascertainment bias due to limited sequence depth

Besides limited sequencing depth also sequence quality is a limiting factor for calling SNPs. This is illustrated by the global trend in the number of predicted SNPs per read position in d-RRL and d-Shared (Figure 2A and 2B) which are mirroring the decreasing trend of average base call score per nucleotide position inherently present in Illumina sequencing (as also observed in our data set, data not shown). A similar trend is not observed in d-Between because here the SNPs are predicted from differences between the reference and the pool wild mallards. Read depth is less limiting in d-Between because the read depth is only used to provide one unambiguous (consensus) base, deviating from the reference, of sufficient quality whereas in d-RRL the read-depth has to provide sufficient calls for both the major allele and the minor allele to be called as a SNP.

Besides the unequal distribution of identified SNPs over the read positions also the underrepresentation of SNPs with a MAC <0.2 is an indicator of a coverage limitation. Due to the limited coverage, only SNPs that are present in multiple individuals in the pool have a reasonable probability to meet the minor allele representation constraint set by our SNP detection method. More common alleles will pass the representation constraint more frequently than rare alleles resulting in an overrepresentation of common alleles and an underrepresentation of rare alleles.

SNP quality assessment by comparison

We identified a large number of putative SNPs in the sequenced pool by sampling ~5% of the mallard genome. Extrapolating the total number (d-RRL + d-Between) of identified SNPs would result in a SNP every ~375bp. The actual number of true SNPs in d-RRL and d-Between is expected to be lower considering the over representation of predicted SNPs in the read positions one to six together with low TS:TV ratios in these read positions. Also the comparison of d-RRL with d-WGS, in which common true variants remained and false SNPs are discarded, show that SNPs predicted in read positions one to six should be used cautiously. The distribution of d-Shared does not show over

representation of SNPs on position one to six. Furthermore, expected TS:TV ratios in d-Shared were observed for positions three to six and expectedly lower TS:TV ratios in position one and two due to the RRL enzyme restriction motif. Therefore we think that a considerable fraction of SNPs in read positions one to six in d-RRL and d-Between are false positives. Because standard sequencing error rates of the Illumina GAII are low (<0.005) in the first 20 bases of a read [29] we expect that the first 6 bases in our sequence dataset were affected by non-standard, systematic, sequencing errors. These are most likely resulting from a combination of inadequate separation of sequencing clusters due to the restriction tag in the RRL and an overloaded sequencing flowcell (Kees-Jan Françoijs personal communication). This hypothesis is sustained by the fact that quality scores were considered by the SNP calling algorithm [28] and that two observations of the minor allele were required for a putative SNP making it unlikely that these numbers of false positives are due to standard sequencing errors. Low TS:TV ratios for SNPs at read position 61 and 62 in d-Between suggest that the SNPs from these positions should also be omitted. Subtracting SNPs from positions one to six (and position 61 and 62 in d-Between) results in 101,095 SNPs in d-RRL and 48,592 SNPs in d-Between that will likely yield good conversion rates in genotyping.

Shared SNPs

We showed that d-RRL shares one sixth of the SNPs with d-WGS and an almost negligible number of SNPs with d-EST. ESTs only represent a few percent of the genome of which only a fraction was sampled by the RRL. Due to this limited shared genome fraction and because SNPs in coding regions are rare, a large overlap in SNPs between these sources was not expected. Between d-WGS and d-EST we observed a relative 2.6 times larger overlap which can be explained by a more or less complete overlap in sampled genome fraction and a better representation of rare alleles in d-WGS. The statistically relatively large overlap between d-WGS and d-RRL indicates a low genetic divergence between wild mallard and domestic duck. A relatively large fraction of shared SNPs between two independent studies is also suggesting a low false discovery rate. As stated earlier, the SNPs identified in this study will be used to study bird migration in a population genetic framework [7]. Because the required number of markers for such an analysis is small compared to the total amount of markers we generated, we selected SNPs from d-Shared that show an equal distribution over the chicken genome. This requirement greatly reduces the number of available markers since only a small fraction could be mapped

(Figure 4) due to the relatively large evolutionary distance between chicken and duck (90 million years)[24]. Genotyping of this SNP subset confirmed the expectation that SNPs which are shared between independent SNP detection studies are of high quality.

Conclusions

When performing SNP identification studies using next generation sequence technologies, it is important to know what limitations in sensitivity and specificity can be expected, particularly at low sequence coverage. We show that sensitivity decreases with decreasing base calling quality towards the ends of sequence reads. which can be compensated for by increasing the sequence coverage in the tailing ends. SNP distribution and TS/TV ratio over read positions are helpful metrics for the assessment of systematic errors in the sequencing dataset in particular when statistics can be compared to a high quality subset of the data. We showed that the fairly large subset of predicted SNPs that is shared between independent SNP detection studies in wild and domestic duck is likely to represent true SNPs, and suggests a low divergence between these subspecies.

Methods

Sample collection and preparation

Mallard DNA samples have been prepared from ethanol preserved whole blood collected from nine individuals from three locations across Europe: two females and a male each from Doñana (Spain), Northern Netherlands and Ottenby (Sweden). Each of these individuals was either directly caught from the wild, or was first generation descendant from local wild mallard parents. DNA extraction was performed using the Gentra Systems Puregene DNA purification Kit according to the manufacturer's instructions. Briefly, ~200µl blood was digested with 9 µg Proteinase K (Sigma) in Cell Lysis Solution (Gentra Systems) at 55ºC over night. Proteins were subsequently precipitated with Protein Precipitation Solution (Gentra Systems) and spun down. DNA from the supernatant was precipitated with isopropanol and washed twice with 70% ethanol. DNA quantity and purity were measured using the Nanodrop ND1000. Possible degradation was inspected on an agarose gel and only high quality DNA samples were used to prepare the DNA pool. Equal amounts of DNA from the nine mallards were combined into two pools of 25 µg each. Aliquots of 5 µg for each pool were digested with either *Alu*I or *Hha*I (10 units per reaction,

Pharmacia). The digested pools in O'range loading dye (Fermentas) were size-fractionated on precast 10% polyacrylamide in 1xTBE with the Criterion™Cell (BioRad). The gel was run 190 minutes at 100 volt and stained for 30 minutes in ethidium bromide solution. After staining, the target fragment size range between 110-130 bp was sliced out of the gel. The gel slice was sheared by nesting a 0.5ml Eppendorf tube (with a hole in the bottom formed with a needle) containing the gel slice inside a 2ml Eppendorf tube, and centrifuged at 14000 rpm for 2 minutes. The sheared gel pieces were covered with 300ul DNA recovery buffer (8mM Tris pH 8.0, 0.08 mM EDTA, 1.25M ammonium acetate), vortexed, and eluted at 4°C overnight, followed by 15 minutes incubation at 65°C. The slurry was divided over two Montage DNA gel extraction devices (Millipore) and centrifuged at 5000g for 10 minutes to purify the eluted gel. DNA was precipitated by adding 1/10 volume 3M sodium acetate pH 5.2, 1 volume isopropanol and 1/500 volume glycogen, washed with ethanol and resuspended in DNA hydration solution (Gentra Systems). The genomic libraries were combined and prepared using the Illumina Sample Preparation kit [31] and sequenced for 76 cycles with the Illumina GAII, Illumina Inc., USA, with a paired end module attached.

SNP detection

Prior to analysis we applied quality filters to the raw reads. Due to the use of restriction enzymes *Alu*I and *Hha*I for creating the genomic libraries we expect that the sequence reads start with a 'C'. Therefore, reads not starting with 'C' were discarded as unreliable or contamination. All reads of the sequencing dataset were trimmed from the position where the average quality score dropped below 12. Reads containing a base that was called with a quality lower than 12 were discarded unless an identical copy of the read occurred in the dataset, since it is unlikely to have two fragments of such a long sequence of nucleotides being identical by chance. We removed reads that - based on the theoretical raw sequencing coverage of the RRL (38X) - were more than four times overrepresented to limit the number of sequences from repetitive regions in the dataset.

As reference genome we used a next generation sequenced domesticated duck reference genome from the Beijing Genome Institute (Huang et al. in prep.). MAQ [28] was used to map the quality filtered reads to the duck genome with default parameters. Putative SNPs were tagged if the reads involved were mapped unambiguously to the reference. We filtered the MAQ [28] SNP output according to several rules: minimal map quality per read: 10; minimal map

quality of the best mapping read on a SNP position: 10; maximum read depth at the SNP position: four times the actual coverage after quality filtering; minimum consensus quality: 10 [20]. We required that the minor allele at a polymorphic position in the reference was observed at least 2 times.

EST-mapping

We mapped d-EST SNPs on the genome reference to identify their genomic locations whereas SNPs in d-RRL and d-WGS were predicted on an identical genome reference coordinate system. Mallard SNPs (with on average 116 bp of flanking sequence) being predicted in EST sequences by the group of Alain Vignal (INRA France, unpublished data) were mapped on the duck reference genome using GMAP [32]. Results were filtered for SNPs that aligned with 96% sequence identity.

Comparative mapping

To examine the distribution of SNPs over the genome, we comparatively mapped our predicted SNPs (including 100bp flanking sequence at each side) on the repeat masked chicken genome. Mapping was performed using BLAT [33] with parameters -oneOff=1 -minIdentity=70.

SNP validation by genotyping

SNPs were validated by genotyping an animal panel using the Illumina GoldenGate® Genotyping assay on an Illumina® BeadXpress with VeraCode™ technology. Selection criteria for the SNPs were based on the Illumina design score (above 0.8) and the assayed 384 SNPs should equally distribute along the chicken genome to minimise the extent of linkage between neighbouring SNPs. Oligo-nucleotides were designed, synthesised, and assembled into oligo pooled assays (OPA) by Illumina Inc. The 384 SNPs were genotyped in 765 animals which included domesticated ducks of a French (7) and a Chinese (189) mapping population, non-*Anas platyrhynchos* (36)_ducks species, ~500 wild ducks from Europe, North America and Asia and the nine ducks that made up the SNP discovery panel. Genotyping results were analysed in Genome Studio (Illumina). The correlation between allele frequency estimated by sequencing and genotyping was calculated over 361 SNP loci that were polymorphic in the discovery panel genotyping by randomly selecting the major or minor allele.

Authors' contributions

HHDK designed and carried out SNP detection, interpreted the genotyping analyses and drafted the manuscript. RHSK collected and prepared DNA samples, performed SNP detection, designed the marker assay, helped in interpreting results and drafting the manuscript. NL and YH provided the BGI reference genome sequence of the domestic duck including the BGI SNPs. AV provided the EST SNPs. MAMG and RPMAC coordinated the research and helped in drafting and revising the manuscript. PvH, JJvdP, JE and HHTP contributed to study design. All authors read, edited and approved the final manuscript.

Acknowledgements

Duck samples for the discovery pool were kindly provided by Jordi Figuerola (Biological Station Doñana, Spain), Marcel Klaassen (NIOO Nieuwersluis, The Netherlands) and (Neus Latorre-Margalef, Ottenby bird observatory and Kalmar University, Sweden). The sources of samples for the genotyping are too numerous to mention, so we thank the enthusiastic wild duck community for their assistance. Technical assistance was provided by Bert Dibbits. We would like to thank Nikkie van Bers for comments on the manuscript, and Hendrik-Jan Megens and Ron Ydenberg for valuable discussions on the subject. This work was financially supported by European Union grant FOOD-CT-2004-506416 (Eadgene), the KNJV (Dutch hunters association), the Dutch ministry of Agriculture, the Faunafonds and the Stichting de Eik trusts (both in The Netherlands). Computational support was offered by the Netherlands National Computing Facilities foundation grant SH-110-08 to RHSK. JE was supported by grant V-220-08 from the Swedish Environment Protection Agency.

References

1. Gilbert M, Chaitaweesub P, Parakamawongsa T, Premashthira S, Tiensin T, Kalpravidh W, Wagner H, Slingenbergh J: **Free-grazing ducks and highly pathogenic avian influenza, Thailand.** *Emerg Infect Dis* 2006, **12**:227-234.
2. Munster VJ, Veen J, Olsen B, Vogel R, Osterhaus ADME, Fouchier RAM: **Towards improved influenza A virus surveillance in migrating birds.** *Vaccine* 2006, **24**:6729-6733.
3. Paul M, Tavornpanich S, Abrial D, Gasqui P, Charras-Garrido M, Thanapongtharm W, Xiao X, Gilbert M, Roger F, Ducrot C: **Anthropogenic factors and the risk of Highly Pathogenic Avian**

Influenza H5N1: prospects from a spatial-based model. *Vet Res* 2009, **41**:28.

4. Atkinson PW, Clark JA, Delany S, Diagana CH, du Feu C, Fiedler W, Fransson T, Gaulthier-Clerc M, Grantham M, Gschweng M: **Urgent preliminary assessment of ornithological data relevant to the spread of Avian Influenza in Europe.** In *Report to the European Comission.* Edited by Delany S, Veen J, Clark A. Wageningen, The Netherlands: *Wetlands International*; 2006.
5. Bauer H, Bezzel E, Fiedler W: **Kompendium der Vögel Mitteleuropas.** Wiebelsheim, Germany: Aula-Verlag; 2005.
6. Scott DA, Rose PM: ***Atlas of Anatidae populations in Africa and Western Eurasia.*** Wageningen, The Netherlands: *Wetlands International* Publication No. 41, Wetlands International; 1996.
7. Wink M: **Use of DNA markers to study bird migration.** *Journal of Ornithology* 2006, **147:**234.
8. Morin PA, Martien KK, Taylor BL: **Assessing statistical power of SNPs for population structure and conservation studies.** *Molecular Ecology Resources* 2009, **9:**66.
9. Ryman N, Palm S, André C, Carvalho GR, Dahlgren TG, Jorde PE, Laikre L, Larsson LC, Palmé A, Ruzzante DE: **Power for detecting genetic divergence: differences between statistical methods and marker loci.** *Mol Ecol* 2006, **15**:2031-2045.
10. **Food and Agriculture Organisation of the United Nations** [http://faostat.fao.org/].
11. Huang C, Cheng Y, Rouvier R, Yang K, Wu C, Huang H, Huang M: **Duck (Anas platyrhynchos) linkage mapping by AFLP fingerprinting.** *Genet Sel Evol* 2009, **41**:28.
12. Huang Y, Zhao Y, Haley CS, Hu S, Hao J, Wu C, Li N: **A genetic and cytogenetic map for the duck (Anas platyrhynchos).** *Genetics* 2006, **173**:287-296.
13. Bennett S: **Solexa Ltd.** *Pharmacogenomics* 2004, **5**:433-438.
14. Bentley DR: **Whole-genome re-sequencing.** *Curr Opin Genet Dev* 2006, **16**:545-552.
15. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G: **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.** *Nucleic Acids Res* 2006, **34**:e22.
16. Altshuler D, Pollara VJ, Cowles CR, Etten WJV, Baldwin J, Linton L, Lander ES: **An SNP map of the human genome generated by reduced**

representation shotgun sequencing. *Nature* 2000, **407**:513-516.
17. Kerstens HHD, Crooijmans RPMA, Veenendaal A, Dibbits BW, Chin-A-Woeng TFC, Dunnen JTD, Groenen MAM: **Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey.** *BMC Genomics* 2009, **10**:479.
18. van Bers NEM, van Oers K, Kerstens HHD, Dibbits BW, Crooijmans RPMA, Visser ME, Groenen MAM: **Genome-wide SNP detection in the great tit Parus major using high throughput sequencing.** *Mol Ecol* 2010, **19 Suppl 1**:89-99.
19. Sánchez CC, Smith TPL, Wiedmann RT, Vallejo RL, Salem M, Yao J, Rexroad CE: **Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library.** *BMC Genomics* 2009, **10**:559.
20. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z, Kerstens HH, Law AS, Megens H, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Tassell CPV, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS One* 2009, **4**:e6524.
21. Wiedmann RT, Smith TPL, Nonneman DJ: **SNP discovery in swine by reduced representation and high throughput pyrosequencing.** *BMC Genet* 2008, **9**:81.
22. van Tassell CPV, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS: **SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.** *Nat Methods* 2008, **5:**247-252.
23. Fillon V, Vignoles M, Crooijmans RPMA, Groenen MAM, Zoorob R, Vignal A: **FISH mapping of 57 BAC clones reveals strong conservation of synteny between Galliformes and Anseriformes.** *Anim Genet* 2007, **38**:303-307.
24. Skinner BM, Robertson LBW, Tempest HG, Langley EJ, Ioannou D, Fowler KE, Crooijmans RPMA, Hall AD, Griffin DK, Völker M: **Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis.** *BMC Genomics* 2009, **10**:357.
25. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
26. Amaral AJ, Megens H, Kerstens HHD, Heuven HCM, Dibbits B, Crooijmans

RPMA, den Dunnen JT, Groenen MAM: **Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome.** *BMC Genomics* 2009, **10**:374.

27. Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD: **Eukaryotic genome size databases.** *Nucleic Acids Res* 2007, **35**:D332-8.
28. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858.
29. Kao WC, Stevens K, Song YS: **BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing.** *Genome Research* 2009, **19:**1884.
30. Scarano E, Iaccarino M, Grippo P, Parisi E: **The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos.** *Proc Natl Acad Sci U S A* 1967, **57**:1394-1400.
31. Illumina: **Protocol for Whole Genome Sequencing using Solexa Technology.** *BioTechniques Protocol Guide* 2006,**12**:29.
32. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21**:1859-1875.
33. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656-664.

6 Discussion

General discussion

In animal breeding the principles of genetics and biometry are applied to exploit the genetic potential of farm animals to improve the efficiency and sustainability of production and to secure food availability. Principles of animal breeding were applied to change animal populations thousands of years before the sciences of genetics and biometry were formally established. In the Introduction of this thesis I already give an impression of how the recent availability of a vast amount of molecular markers can contribute to significant genetic gains in breeding and increase our knowledge of genome structure and evolution.

The aim of this thesis is to contribute to the development of a genetic variability repository for farm animals. We designed and validated DNA sequence data analysis pipelines for obtaining genetic markers in farm animals which can be used in the construction of linkage maps, SNP genotyping based estimation of kinship and pedigree reconstruction and QTL studies. We focused on bioinformatics methods to analyze whole genome sequencing data from both traditional capillary and next generation DNA sequencing (NGS) platforms. We targeted for the development of a cost effective and reliable prediction method to identify thousands of true single nucleotide polymorphisms (SNPs) in species with and without a sequenced reference genome. We also aimed for a method to identify structural variants (SVs) at a high resolution in an available NGS data set. This dataset was created on a reduced representation of the chicken genome and allowed getting a first glimpse of the abundance of structural variation in an avian genome. In this final chapter I will evaluate the design and implementations of the constructed data analysis pipelines in the light of the recent revolution in DNA sequencing. I also show the results of our data analysis pipelines and to what applications our developed marker repositories can contribute to (Table 1). Finally I will discuss the current status and future perspectives in the discovery of genetic variation in livestock.

SNP mining in publicly available sequence repositories

In Chapter 2 we describe the design of an approach to identify thousands of genetic markers in the pig genome using public sequencing databases. At the time this work was performed, the pig genome sequencing project [1] had been running for two years and the first release of the pig genome reference sequence was expected to take at least another two years.

Table 1: Applications that our developed markers sets can contribute to

Species	Class of genetic variation	Applied technique for identification	Nr of obtained markers	Applications for detected markers
Chicken	SV	Paired-end sequencing of genomic RRLs and mapping to reference genome	188	Development of a repository of SVs in chicken for future SV genotyping based association studies (currently not feasible)
Duck	SNP	(Paired-end) sequencing of genomic RRLs and mapping to reference genome	149,687	Improving the duck genome markers and linkage maps allowing the chromosomal assignment of the sequence scaffolds of the sequenced reference genome. These dense maps will also improve resolution in QTL studies and allow for duck migration studies in a population genetic framework
Pig	SNP	Mining public sequence repositories and mapping to BAC and BAC-end sequences	6,374	Genotyping based estimation of kinship or pedigree reconstruction and QTL studies.
Turkey	SNP	Sequencing of a randomly sheared genomic RRLs and mapping to assembled contigs	11,287	Improving the turkey genome markers and linkage maps which aids in the chromosomal assignment of the sequence scaffolds of the sequenced reference genome. These improved maps also allow for QTL detection at higher resolution.

RRL = reduced representation library
BAC = bacterial artificial chromosome
QTL = quantitative trait loci

Within the framework of this sequencing initiative, pig bacterial artificial chromosomes (BACs) that were completely or partially sequenced and BAC end sequences including their naming and mapping information were made available to researchers worldwide. Furthermore, a previous pig sequencing initiative on five different breeds had resulted in 4.8 million whole genome shotgun sequence reads of at least 150 bp stored in a public database and providing a genome coverage of 0.66 [2].
Assuming that the whole genome shotgun sequence reads are not equally represented, this dataset [2] potentially holds the information for a large number of Single Nucleotide Polymorphisms (SNPs). Our highly automated SNP identification pipeline, consisting of publicly available softwares, enabled the identification of thousands of high quality SNPs. The use of a sequence alignment program in combination with a computational method to identify polymorphisms, is a common method for automated de novo sequence-based SNP detection and has been described for SNP detection in ESTs [3-5]. However a genomic implementation by the integration of BAC mapping and naming information in the pipeline to compensate for a lack of a reference genome sequence is unique in its kind. Read mapping performance was facilitated by including mate pair information of shotgun sequences. The pipeline was implemented on a computer cluster, which enabled efficient mining of large sequence data sets in parallel.
A total of 98 thousand SNPs were identified of which 6,374 had a redundancy of two or more. This work showed that mining of public whole genome shotgun sequence databases can result in the identification of thousands of SNPs. The identified SNPs distribute equally over the genome and have a performance that is comparable to those available in existing SNP repositories. Our SNP detection pipeline has resulted in an increased SNP marker density on the pig genome. One intended applications of these identified SNPs was the development of a marker based method for the estimation of kinship and pedigree reconstruction in pig. For this purpose we provided a, at that time considerable, set of thousands of markers. Recently hundreds of thousands of markers became available in pig [6] due to the availability of the draft genome sequence of the pig and the launch of NGS technology. NGS allows for SNP detection over a considerable larger part of the pig genome, in larger number of individuals and breeds, at affordable costs.
The developed SNP detection pipeline consists of universal pieces of software and queries common public data repositories. Nevertheless its applicability in other species is probably limited by its specific requirements. For at least part of

the mapped BACs, BAC sequences have to be available for clustering whole genome shotgun reads by sequence homology. Because only clustered reads homologous to a mapped BAC sequence can lead to SNPs that can be assigned to a genomic location by using the BAC mapping information. Furthermore, a sufficiently large whole genome sequence repository derived from multiple individuals is a prerequisite to identify SNPs by sequence alignment. Most animal genome sequencing projects, however, are primarily targeting for a whole genome sequence of one individual and perform only low coverage re-sequencing on a few individuals afterwards.

A future approach to SNP mining in public data

The implementation of automated SNP identification pipelines will become challenging in the near future. Until recently, these pipelines were handling data coming from a single sequencing platform which has been the sequencing standard for two decades. Nowadays multiple sequencing platforms with different types of raw data output have emerged. Although the platforms Illumina GAII and the Helicos HeliScope approximate the traditional capillary sequencers, the platforms Roche GS FLX and SOLiD produce a completely different raw sequencing signal and data conversions will be required to compare or merge sequencing data from different platforms. However, data conversion will introduce uncertainties and loss of important information to e.g. discriminate between sequencing error and polymorphism. To, reliably, compare sequences coming from a variety of platforms, new base calling algorithms need to be developed that convert platform-specific quality scores and sources of errors into a universal quality scale. Such information is needed in the sequence alignment stage of the pipeline. Short reads with, compared to traditional Sanger sequencing, substantial error rates can be aligned more precisely by including quality scores in the alignment of these sequence reads [7]. This will avoid misinterpretation of misaligned reads as positional variants. Finally, the sequence quality and platform specific biases will determine how often a sequence variant has to be detected to be scored as a true polymorphism. Therefore, the development of algorithms for automated SNP detection on sequencing data coming from different platforms will remain a big challenge. The enormous variation in platform biases and error rates is a big hurdle for implementing an all in one sequence analysis solution.

Given the recent developments in whole genome sequencing, another important issue related to the implementation of a SNP-detection pipeline as described in Chapter 2 is the requirement for a high quality BAC map. The more than 1000-

fold increase in sequencing throughput has reduced the time span between the moment that a BAC map becomes available and the moment that a reference genome sequence is released. Implementing a SNP detection pipeline that roughly maps predicted SNPs on a BAC-map is most likely not worth the effort when SNPs can be mapped with a base pair precision on the reference sequence just a few months later. In addition it is questionable whether upcoming NGS genome sequencing projects will invest in generating high quality BAC maps prior to genome sequencing. Sequence length, sequence quality and coverage are all increasing and, together with the development of mate pair sequencing on the NGS platforms, provide an increased texture to the sequence data. These developments will make *de novo* assembly using only the sequence data feasible in the near future. However, for a complete, true and unbiased *de novo* assembly of complex genomes the availability of a BAC map is still extremely useful, although for most genome sequencing projects (less laborious) comparative mapping approaches are being used [8]. Moreover, the recent developments in DNA sequencing have reduced sequencing cost drastically and provide, in particular if combined with a sequencing target complexity reduction, an attractive alternative for BAC mapping. Such alternatives are described in Chapters 3 and 5 of this thesis and will be further discussed in the next paragraph.

Mining for SNPs in unsequenced genomes by next-generation sequencing

Cost effective sequencing and reconstruction of a large complex animal genome is currently still a major challenge. However, reduced complexity representations of genomes are ideal sequencing targets for SNP detection as is shown in Chapters 3 and 5 of this thesis. The recent improvements in read length and sequence quality and the decreased cost per base, facilitate obtaining sufficient coverage of the fragments present in a genome library. For the approach described in Chapter 3, random shearing of the genome library was a necessity to sufficiently cover the fragments of the library by short sequence reads. For the approach used in Chapter 5, the read length and the mate pair information was sufficient to cover the (considerably smaller) library fragments. Sequenced and assembled library fragments subsequently served as reference subgenome providing sufficient sequence context to identify and map SNPs. We have shown that the short reads resulting from a few NGS runs on reduced representation libraries provide both genomic context and ample coverage to perform sequence variant discovery. Thousands of high quality

SNPs were obtained of which the turkey SNPs showed a genotyping conversion rate of 95%. Due to the lack of a complete genome build in turkey and duck, we used a comparative mapping approach to assign predicted SNPs to a genomic location. The prediction of the genomic locations of SNPs in the target genome by mapping the sequenced library fragments to the chicken chromosomes, might have introduced some bias towards the physical map of the reference species. However, we show for turkey that it is an effective approach for the selection of SNPs to be included in genotyping assays for constructing the linkage map. The successful construction of a turkey linkage map, which assisted in the turkey reference genome build, shows that a comparative approach is effective in reconstructing genomes from NGS sequence contigs and scaffolds (Turkey genome paper submitted). From the chicken-turkey comparative map we can conclude that chicken and turkey karyotypes (common ancestor ~28 MYA) have undergone relatively very few chromosomal rearrangements during evolution, which is consistent with an earlier study [9]. A similar approach is currently being performed to build the duck reference genome. Constructing the chicken-duck comparative map requires duck SNPs that show an even distribution on the chicken reference genome. We observed that for duck, compared to turkey, a considerable lower fraction of SNPs could be mapped. Nevertheless, based on thousands of SNPs that still could be mapped and sufficiently cover the chicken genome, I expect that these will allow for duck linkage map construction and facilitate the duck reference genome build. This would mean that even evolution periods of ~90 MYA can be spanned in avian genome assembly using comparative genomics approaches. The numbers of identified markers will not allow for genome wide association studies (GWAS) in turkey because recent linkage studies in the highly similar chicken genome [10] suggest that at least 100,000 SNPs are required to exploit haplotype information. However the availability of a complete turkey reference genome allowed for exploiting the full potential of the nucleotide variation in our sequencing dataset. Using this reference we revealed a total of 37,041 putative SNPs which is a step closer towards the 100k SNPs required for GWAS in turkey. The sum of SNPs identified by our analysis (Chapter 5) within the wild duck pool and between wild duck and domesticated duck already exceeds 100,000. Together with the high quality SNPs identified in whole genome sequencing and EST sequencing in duck the here identified SNPs will very likely allow for GWAS in duck.

Further improvements

Although increasing read length and mate pair information already allows for complete sequence coverage of library fragment sizes of a few hundred base pairs, larger insert sizes still require that libraries have to be randomly sheared, sequenced and reconstructed by sequence assembly. Equal sequence coverage of the large fragment libraries facilitates sequence assembly but is not obtained without taking measures in sample preparation (Chapter 3 and [11]). Recently developed sample preparation techniques avoid overrepresentation of sequences derived from the ends of the fragment of the library and maximize coverage yield and limit coverage variability [11]. I believe that these preparation techniques will contribute to the reconstruction of a larger fraction of the reduced representation library (RRL) and that it would have increased the number of reconstructed library fragments that could be mapped in our turkey study (Chapter 3). This assumption is supported by a recent publication describing a "new" strategy for whole genome sequencing by partitioning the genome using RRLs prior to assembly [12]. The authors state that advances in sequencing technology and approaches will facilitate the sequencing of RRLs containing more fragments with fewer reads. This will allow for whole genome sequencing of mammalian-sized genomes using a relatively small number of fragment pools. This raises the question of how long unsequenced genomes will remain unsequenced.

How long will unsequenced genomes remain unsequenced?

The human, mouse, cattle, chicken, fruit fly and many microbial species are examples of groups of species for which a sequenced reference genome is available. I have experienced that the number of species within this group is growing exponentially as a result of the increasing output at decreasing cost of NGS sequencers. As an example: within one year, a few months after submission of Chapter 3 to a scientific journal and in the course of the analysis for Chapter 5, next generation sequenced genomes for turkey and duck became available for our group but not yet in public domain. Initially NGS has been applied for (re-)sequencing of relatively small sequencing targets like complete prokaryote genomes and eukaryote genomes for which the sequencing target had been reduced e.g. by using individual BACs. More recently data from a NGS whole genome shotgun approach has been used to close small gaps in the assembly of traditional Sanger reads and to assist in scaffolding the genome by providing mate pair data as a result of paired end sequencing of genomic libraries with specific large inserts [13]. A next generation genome scale *de*

novo sequencing approach has been reported for bacteria [14-16] but typical results in thousands to millions of contigs and scaffolds when applied on much larger and more complex sequencing targets, like the duck (Huang et al. in preparation), turkey (Dalloul et al. submitted), giant panda [17], cucumber [13] and human (Table 2) genomes. Reconstruction of repetitive genomic element using NGS is hard because repeat sequences prevent short reads from being assembled unambiguously. Furthermore each NGS platform produces a unique reproducible pattern of variable sequence coverage [19]. By mixing of different NGS read types in the assembly regions of low coverage in one NGS approach might be sufficiently covered by the other. This theory is supported by mixing Roche/454 and Illumina read data which indeed resulted in improved *de novo* assemblies of microbial genomes compared with assemblies based on data from either platform alone [20,21]. Promising results in whole genome sequencing by partitioning using RRLs and sequencing of these RRLs by a mix of NGS read types will likely make *de novo* assembly of mammalian-sized genomes soon feasible.

Table 2: Achievements of current de novo short fragment sequence assemblers

assembler	*n*	*contig N50*	*scaffold N50*	*genome coverage*	*species*	*Short read archive*	*reference*
SOAPdenovo		7.4kb	446.3kb	87.4%	human (Asian)	ERA000005	[8]
SOAPdenovo		5.9kb	61.9kb	85.4%	human (African)	SRA000271[2]	[8]
ABBYS	2.76 M	1499 bp		69%	human (African)	SRA000271	[18]
SOAPdenovo		4611bp		85%	human (African)	SRA000271	[8]
SOAPdenovo		40kb	1.3Mb		giant panda		[17]
SOAPdenovo		12.5	172kb		cucumber		[13]

[1]*contigs equal to or larger than 100 bp*
[2]*supplemented with large insert library*

A first glimpse of the extent of structural variation in an avian genome

SV markers are rarely used in animal genetics but might gain interest once their abundance is estimated and their putative relation with phenotypic differences is understood. Do SVs influence phenotypes and what is the impact of SVs on animal genome evolution? Previously used methods in CNV screening like fluorescent in-situ hybridization (FISH) of bacterial artificial chromosomes (BACs) and array based comparative genome hybridisation (array-CGH) have a limited resolution. Due to this limit in resolution the majority of the identified SVs have not yet been finely resolved to the nucleotide level. For most reported CNVs in animals, we do not know their true population frequency because they have not been genotyped. To be able to study the potential role of SVs in phenotypes and in genome evolution, we will need a more complete catalog of SVs in animal genomes and large SV genotype datasets from different populations. Precise *de novo* CNV mutation rates throughout the genome are required to better understand the contribution of CNV versus SNP to genome evolution, particularly with respect to gene duplication/triplication and exon shuffling [22].

Low requirements for obtaining a promising first glimpse

In Chapter 4 we provide a first glimpse of the extent of structural variation in an avian genome at a ~50 bp resolution. We show that even the analysis of mate pair information of a paired end sequenced reduced representation library is sufficient to predict several hundreds of candidate structural variants (SVs) in the chicken genome. More than 180 of these SVs are very likely to represent true structural variation between four chicken breeds and red jungle fowl. The sequencing and the bioinformatics approach we used put high constraints on the SV detection thereby putatively ignoring true variants. Future validation studies can be considered to find out what constraints can be relaxed, at tolerable false positive rates, to increase sensitivity of the detection method. The majority of SVs identified by our method were small deletions, which is consistent with an earlier study where an inverse relationship between the number of SVs in the human genome and their size was established [23]. Our detection strategy did allow for the detection of insertions in only a very limited size range (few tens of basepairs). I expect that the actual size and frequency of the total number of small insertions in the chicken genome is similar to the observed number of small deletions. Based on our findings in Chapter 4 we expect thousands of

rearrangements smaller than one kb and hundreds of larger rearrangements in the chicken genome. Furthermore our study identified SVs in coding regions of the genome suggesting that some of the small SVs putatively can be related to phenotypes. Based on this first glimpse, I think, there is evidence that SVs considerably contribute to phenotypes and genome evolution and that it is worthwhile to obtain a more complete picture of the extent of this type of genetic variation in animal genomes.

More demanding approaches completing the catalog of SVs

Our SV detection method can be classified as a paired-end resequencing and mapping approach using standard insert libraries. Paired-end mapping approaches combined with high-throughput sequencing [23-28], (Chapter 4 of this thesis) provide the possibility of reliably detecting SVs that are one to three orders of magnitude smaller than those assayed previously using FISH mapping of BACs or array-CGH or lower-density oligonucleotide arrays. Most recent advances in paired-end sequencing, which were not available at the time of our study, are the use of large insert library kits and increased read lengths. The latter will improve mapping accuracy whereas the first will allow for deep paired end sequencing of insert libraries in a larger size range.

Paired-end reads of large insert libraries will allow for spanning repetitive elements, which likely hold the majority of genomic structural variation [29-31]. By constructing libraries from a randomly sheared genome, each SV will be predicted by paired end reads from a variety of genome fragments sampled from that genomic region. This will facilitate breakpoint resolution and reduces required additional PCR and sequencing efforts [32]. In spite of these improvements, the SV detection strategy by paired end sequencing and mapping (PEM) still has a fairly large false negative detection rate for large structural variants and segmental duplications compared to more laborious techniques such as fosmid paired end sequencing (FPES) or oligonucleotide arrays comprising millions of optimized probes [33]. Segmental duplications are more difficult to ascertain using PEM because many of the reads in these regions do not map to unique locations in the genome [24]. However, a quantitative NGS approach for detecting segmental duplications can be used to complement the paired -end mapping technique. In this approach the depth of coverage in sequence data is analyzed to look for genomic regions that differ in copy number between individuals [34].

A shortcoming of reference based SV detection techniques like PEM and array-CGH is the bias towards the reference genome. In a sequencing context, reads

obtained from large genomic regions that are missing in the reference cannot be mapped whereas in a micro-array context these genomic regions are not represented by probes. This lack of genomic information will potentially hide structural variation between the sampled individuals. Therefore, the most versatile strategy for SV detection is sequencing and unbiased *de novo* assembly of individual genomes [8]. This approach will undoubtedly result in a more accurate and complete catalog of structural variation in a genome. However it is unclear what sampling depth is needed to reliably capture the majority of SVs. In human there is evidence that there are many SVs related to disease present within the general population with frequencies lower than the classical definition of a polymorphism (>1%)[35].
Furthermore a linear representation of a genome which is currently being used is not proper to capture and represent all structural variation and therefore needs to be replaced by a higher level of data storage and visualization. Because of the costs of whole genome sequencing and the impossibility to reconstruct complex genomes by *de nov*o assembly of NGS data this approach of SV detection will remain unfeasible in animal sciences for the near future. Even if we had (almost) completed the catalog of structural variation, it would not be possible to genotype CNVs genome-wide due to the lack of a robust manner. Currently the degree of uncertainty in genotype inference reduces the power of association studies, and potentially increases the risk of false-positive associations.

SVs and (unraveling) their relation to phenotype

The biological effect and the evolutionary process behind medium sized (10-50 kb) and small (<10 kb) SVs, which are thought to represent the majority of SVs in the human genome, remains currently largely unknown. These SVs generally have been below the reliable detection limit, and thus are underrepresented in current databases [35,36].
For large SVs, studies in human have provided evidence for their involvement in gene regulation by various molecular mechanisms, including gene dosage, gene disruption, gene fusion and position effects. A well known example of the influence of SVs on phenotype is the deletion of the alpha-globin gene resulting in alpha-thalassaemia in homozygous carriers [37] and protection against malaria in heterozygous carriers [38]. Altered regulation caused by SVs has also been associated with Mendelian [39,40] as well as sporadic traits, and also has been associated with complex diseases like Parkinson disease, Alzheimer disease, mental retardation, Autism and Schizophrenia in human. Furthermore

recent studies have reported that altered expression levels due to CNVs affect susceptibility to HIV, Crohn disease, psoriasis, Pancreatitis, Systemic lupus erythematosus and glomerulonephritis [41]. Moreover, CNVs can also represent benign polymorphic variants and in particular gene duplication and exon shuffling are thought to be a predominant mechanism driving gene and genome evolution. As stated in the Introduction of this thesis only a limited number of animal traits have been linked to CNVs. Currently a large scale CNV detection study is being performed at an eight kb resolution using array based comparative genome hybridization (R. Crooijmans, personal communication). The study of CNVs, in particular those that result in gene amplification favored by positive selection, may reveal genomic regions that were evolutionally favored because of their adaptive benefits. Genomic alteration due to major environmental impact (e.g. domestication) can be identified and modified genomic regions might be linked to traits or hide thus far undiscovered functional genes.

Common SVs in human seem to show patterns of allele frequency, linkage disequilibrium and population differentiation that mirror the properties of SNPs [42]. Cataloging the genomic locations, haplotypes and sequence properties of these alternative structural alleles will therefore also be an important direction for completing databases of common patterns of genetic variation in animals. A complete catalog encompassing SNPs and SVs can be used when attempting to unravel the molecular genetic basis of a given phenotype. In other words SNP based linkage and association studies should be supplemented by SV based linkage and association studies. Traits previously intractable by conventional genetic (SNP) analysis may become manageable by including SVs in the analysis, as was shown for autism spectrum disorders in human [43]. Furthermore the simultaneous study of SNPs and SVs, both common and rare, will be needed to understand the relative contribution of each form of variation to traits in animal populations.

Discovery of genetic variation in animals, what can be expected in the near future

Discovery of genetic variation in the near future needs to include both the identification of variation at the nucleotide level and the profiling of structural variation between the reference and individuals. Ultimately this is accomplished by *de novo* assemblies allowing a sound identification of DNA sequence information that is unique to each population. More sources of genetic variation can be captured by analysis of individual genomes, transcriptomes and epigenetics. Collectively, these data will facilitate a more predictive biological approach to the study of phenotypes in farm animals and provide more understanding of the genetic basis of traits. Acquired understanding of the genetic basis of traits can be applied in breeding programs to improve product quality, production efficiency and to reduce the incidence and impact of disease.

Individual genomes

NGS of genomes enables the genotype-phenotype correlations to be studied in the context of the whole genome. The potential utility of individual sequencing has been illustrated by a recent cancer study in which complete genomes of healthy and affected tissues were sequenced and intra-individual genome comparisons revealed cancer related somatic mutations [44,45]. Currently sequencing based association studies reveal genotypes by evaluating the number of reads obtained from specific chromosomes and thus can replace map based genetics [46].

To fully understand genome function and evolution, the complete sequence of multiple individuals representing a population of a species will be required. In humans such sequencing projects are ongoing and have already resulted in the publication of the complete genomes of five individuals [25,27,28,47,48]. For human and mouse high-resolution profiles of genomic variation will soon be available(www.1000genomes.org,www.sanger.ac.uk/modelorgs/mousegenomes). NGS are revolutionizing these sequence gathering efforts and help to obtain whole genome sequence for additional species of interest. With the advent of ultra low cost sequencing technology, routine sequencing of individual animal genomes will become real within a period of 5-10 years. Using this data it will be possible to determine precisely which regions in animal genomes are actually functional, allowing variants found in those regions to be prioritized for follow-up. The goal is to characterize the genetic architecture of animal diseases and complex traits, moving beyond the common SNPs that are

currently forming the backbone of genome-wide association studies.

Transcriptome sequencing

Transcriptome sequencing is a reduced target sequencing approach and has been used for applications ranging from gene expression profiling, non coding RNA discovery and quantification, genome annotation to rearrangement detection. A next-generation high-throughput transcriptome sequencing approach has as a unique feature that the data can be analyzed in various ways. It can be analyzed to provide insight into the level of gene expression, the structure of genomic loci can be analyzed, and sequence variation present at loci (e.g., SNPs) or due to RNA editing can be detected. In animal sciences NGS will economize identification and quantification of mRNA and small regulatory RNAs under different conditions or in different cell types. Thereby it may replace micro-array based techniques [49], serial analysis of gene expression (SAGE)[50] and massively parallel signature sequencing (MPSS) [51]. Major benefits are the less stringent standardization and replication requirements and a more robust detection of rare RNAs provided by high sampling depth.

The increasing number of sequenced genomes of animal species fueled by the advances in sequencing technologies has emphasized that techniques for the annotation of protein-coding genes involving the elucidation of their correct exon-intron structures are lagging behind. EST driven techniques, currently being the standard for protein coding gene annotation, fail to cover 20-40% of transcripts including rare transcripts, transcripts with highly specific expression patterns, extremely long transcripts, transcripts as a result of alternative splicing and transcripts coming from complex loci [52]. Next generation sequencing has the potential to capture rare transcripts by providing much deeper coverage of EST libraries. Paired end sequencing technology and increased read length will allow for a cost effective detection of a higher variety of transcripts in the near future. This has already been illustrated in a cancer study in human in which rearrangements in the transcribed part of the genome were detected by a paired-end ditag transcriptome sequencing methodology [53].

In animal sciences transcriptome sequencing is a valuable technique to complete gene catalogs in animals and to quantify gene expression levels. A complete catalog of genes and their expression levels measured in animals showing trait differences will be helpful to correlate genetic variation in genes or variation in expression level to these traits.

Next-generation sequencing and epigenetics

Epigenetics is the study of heritable gene regulation caused by other mechanisms than the DNA sequence itself. The two major types of epigenetic modifications regulating gene expression are DNA methylation by covalent modification of cytosine-5′ and posttranslational modifications of the amino acids that make up histone proteins. Methylated DNA regions tend to be transcriptionally less active, through a mechanism not fully understood whereas modified histone proteins might change the way that DNA is wrapped around nucleosomes which influences gene expression as well. Important specific epigenetic processes in animal breeding include imprinting, gene silencing and maternal effects. Epigenetics also accounts for some of the mechanisms explaining why differentiated cells in a multicellular organism are programmed to only express the genes that are necessary for their own activity. Studies in human have implicated that epigenetic modifications are of prime importance in oncogenesis and development, setting the grounds for the Human Epigenome Project [54] and the forthcoming Cancer Epigenome initiative [55].

Bisulphite sequencing is one of the approaches used to detect DNA methylation and currently next-generation sequencing makes it feasible to perform genome-scale bisulphite sequencing on large-mammalian genomes using reduced representation libraries and thus is providing a scalable and valuable tool for epigenetic profiling of cell populations [56]. Posttranslational covalent modification of histone tails including methylation, acetylation, phosphorilation and ADP ribosylation are thought to have an effect on the strength of DNA-histone interactions thereby determining the accessibility of DNA to transcriptional regulators [57]. Histone modifications have been identified by chromatin immunoprecipitation (ChIP). NGS is currently being used for the high throughput sequence based characterization of bound DNA (ChIP-Seq). ChIP–seq offers higher resolution and cleaner data at a lower cost than the array-based alternatives for genome wide profiling of large genomes and will allow the profiling of a large number of DNA binding proteins, as well as a more complete set of chromatin marks in thousands of epigenomes across multiple tissues, cell types, conditions and developmental stages [58].

Epigenetic mechanisms and their effects in gene activation and inactivation, are increasingly understood to play a considerable role in phenotype transmission and development. To increase the understanding of this role in animal reproductive biology and to understand how mammalian species are regulated by imprinting, next generation sequencing can be used to perform a comprehensive analysis of imprinted genes in animals.

Now we can efficiently generate data for variant discovery; how we gonna analyze and store it?

NGS technology promises to deliver cost effective genome coverage in many applications in the very near future, allowing individual genome sequences, comprehensive transcriptome sequencing and epigenetics. Using this information we can generate more complete reference genomes with fewer gaps in which SNP and SV information is integrated. A comprehensive reference genome including the genomic variation observed in many individuals representing populations will not fit in the currently used linear representation. Most likely the next generation reference genome will consist of data structures and compression algorithms holding the multi dimensional complex of genomic sequence data.

Therefore, some hurdles in software development and solving hardware deficiencies have to be taken to efficiently handle and analyze these amounts of high complexity data and to keep them accessible.

Great challenges for many laboratories are likely to be the effective management and analysis of the immense amount of sequencing data in order to make NGS applications a routine. This will require the development of efficient and robust software tools and pipelines for data analysis. We have shown that close interaction between geneticist, laboratory scientists and bioinformaticians is fruitful to take up the gage of handling and interpreting this data.

References

1. Schook LB, Beever JE, Rogers J, Humphray S, Archibald A, Chardon P, Milan D, Rohrer G, Eversole K: Swine Genome Sequencing Consortium (SGSC): **A Strategic Roadmap for Sequencing The Pig Genome.** *Comp Funct Genomics* 2005, **6**:251-255.
2. Wernersson R, Schierup MH, Jørgensen FG, Gorodkin J, Panitz F, Staerfeldt H, Christensen OF, Mailund T, Hornshøj H, Klein A, Wang J, Liu B, Hu S, Dong W, Li W, Wong GKS, Yu J, Wang J, Bendixen C, Fredholm M, Brunak S, Yang H, Bolund L: **Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing.** *BMC Genomics* 2005, **6**:70.
3. Barker G, Batley J, Sullivan HO, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**:421-422.
4. Tang J, Vosman B, Voorrips RE, Linden CGVD, Leunissen JAM: **QualitySNP: a pipeline for detecting single nucleotide polymorphisms**

and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* 2006, **7**:438.

5. Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, Cirera S, Thomsen B, Madsen LB, Høj A, Vingborg RK, Zahn B, Wang X, Wang X, Wernersson R, Jørgensen CB, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Brunak S, Fredholm M, Bendixen C: **SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation.** *Bioinformatics* 2007, **23**:i387-i391.
6. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z, Kerstens HH, Law AS, Megens H, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Tassell CPV, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS One* 2009, **4**:e6524.
7. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858.
8. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20**:265-272.
9. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RPMA, Groenen MAM, Deryusheva S, Gaginskaya E, Carré W, Waddington D, Talbot R, Völker M, Masabanda JS, Burt DW: **Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution.** *BMC Genomics* 2008, **9**:168.
10. Megens H, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH, Hanotte O, Groenen MAM: **Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken.** *BMC Genet* 2009, **10**:86.
11. Harismendy O, Frazer K: **Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology.** *Biotechniques* 2009, **46**:229-231.

12. Young AL, Abaan HO, Zerbino D, Mullikin JC, Birney E, Margulies EH: **A new strategy for genome assembly using short sequence reads and reduced representation libraries.** *Genome Res* 2010, **20**:249-256.
13. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan, Wu Z, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim J, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S: **The genome of the cucumber, Cucumis sativus L.** *Nat Genet* 2009, **41**:1275-1281.
14. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J: **De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18**:802-809.
15. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821-829.
16. Warren RL, Sutton GG, Jones SJM, Holt RA: **Assembling millions of short DNA sequences using SSAKE.** *Bioinformatics* 2007, **23**:500-501.
17. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T, Yiu S, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463**:311-317.
18. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I:

ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009, **19**:1117-1123.

19. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biol* 2009, **10**:R32.
20. Aury J, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P: **High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.** *BMC Genomics* 2008, **9**:603.
21. Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL: **De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae.** *Genome Res* 2009, **19**:294-305.
22. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution.** *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.
23. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318**:420-426.
24. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37**:727-732.
25. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA,

Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Cooley RN, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fajardo KVF, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang G, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Bacigalupo MCR, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JES, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-59.

26. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.

27. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L,

Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-876.

28. Kim J, Ju YS, Park H, Kim S, Lee S, Yi J, Mudge J, Miller NA, Hong D, Bell CJ, Kim H, Chung I, Lee W, Lee J, Seo S, Yun J, Woo HN, Lee H, Suh D, Lee S, Kim H, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang K, Park W, Kim H, Church GM, Lee C, Kingsmore SF, Seo J: **A highly annotated whole-genome sequence of a Korean individual.** *Nature* 2009, **460**:1011-1015.
29. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80**:91-104.
30. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-454.
31. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77**:78-88.
32. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, Vega FMDL, Blanchard AP: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base**

encoding. *Genome Res* 2009, **19**:1527-1541.
33. Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, Eichler EE: **Optimal design of oligonucleotide microarrays for measurement of DNA copy-number.** *Hum Mol Genet* 2007, **16**:2770-2779.
34. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: **Sensitive and accurate detection of copy number variants using read depth of coverage.** *Genome Res* 2009, **19**:1586-1592.
35. Sharp AJ: **Emerging themes and new challenges in defining the role of structural variation in human disease.** *Hum Mutat* 2009, **30**:135-144.
36. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PAW, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40**:722-729.
37. Higgs DR, Pressley L, Old JM, Hunt DM, Clegg JB, Weatherall DJ, Serjeant GR: **Negro alpha-thalassaemia is caused by deletion of a single alpha-globin gene.** *Lancet* 1979, **2**:272-276.
38. Flint J, Hill AV, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, Bhatia K, Alpers MP, Boyce AJ, et al.: **High frequencies of alpha-thalassaemia are the result of natural selection by malaria.** *Nature* 1986, **321**:744-750.
39. Chance PF, Pleasure D: **Charcot-Marie-Tooth syndrome.** *Arch Neurol* 1993, **50**:1180-1184.
40. Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, Ledbetter DH, Greenberg F, Patel PI: **Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A.** *Nat Genet* 1992, **1**:29-33.
41. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution.** *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.
42. Consortium IH: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299-1320.
43. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King M, Skuse D, Geschwind DH, Gilliam TC, Ye K,

Wigler M: **Strong association of de novo copy number mutations with autism.** *Science* 2007, **316**:445-449.

44. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456**:66-72.
45. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, Fulton LA, Locke DP, Magrini VJ, Abbott RM, Vickery TL, Reed JS, Robinson JS, Wylie T, Smith SM, Carmichael L, Eldred JM, Harris CC, Walker J, Peck JB, Du F, Dukes AF, Sanderson GE, Brummett AM, Clark E, McMichael JF, Meyer RJ, Schindler JK, Pohl CS, Wallis JW, Shi X, Lin L, Schmidt H, Tang Y, Haipek C, Wiechert ME, Ivy JV, Kalicki J, Elliott G, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson MA, Baty J, Heath S, Shannon WD, Nagarajan R, Link DC, Walter MJ, Graubert TA, DiPersio JF, Wilson RK, Ley TJ: **Recurring mutations found by sequencing an acute myeloid leukemia genome.** *N Engl J Med* 2009, **361**:1058-1066.
46. Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, Leung TY, Foo CHF, Xie B, Tsui NBY, Lun FMF, Zee BCY, Lau TK, Cantor CR, Lo YMD: **Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma.** *Proc Natl Acad Sci U S A* 2008, **105**:20458-20463.
47. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers Y, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**:e254.
48. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J,

Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456**:60-65.

49. Ness SA: **Microarray analysis: basic strategies for successful experiments.** *Mol Biotechnol* 2007, **36**:205-219.
50. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.
51. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* 2000, **18**:630-634.
52. Brent MR: **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nat Rev Genet* 2008, **9**:62-73.
53. Ruan Y, Ooi HS, Choo SW, Chiu KP, Zhao XD, Srinivasan KG, Yao F, Choo CY, Liu J, Ariyaratne P, Bin WGW, Kuznetsov VA, Shahab A, Sung W, Bourque G, Palanisamy N, Wei C: **Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs).** *Genome Res* 2007, **17**:828-838.
54. Brena RM, Huang TH, Plass C: T**oward a human epigenome.** *Nat Genet* 2006, **38**:1359-1360.
55. Richon VM: **A new path to the cancer epigenome.** *Nat Biotechnol* 2008, **26**:655-656.
56. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**:766-770.
57. Schones DE, Zhao K: **Genome-wide approaches to studying chromatin modifications.** *Nat Rev Genet* 2008, **9**:179-191.
58. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nat Rev Genet* 2009, **10**:669-680.

Summary

Current genetic marker repositories are not sufficient or even are completely lacking for most farm animals. However, genetic markers are essential for the development of a research tool facilitating discovery of genetic factors that contribute to susceptibility to disease, to protection against illness and to overall welfare and performance in farm animals.

Genomic analyses of related species reveals the evolution of genome organization as well as supports the identification of the genetic background of economically and biologically important traits. The availability of genetic linkage maps and genome sequence information that is conserved across compared species, enables the construction of comparative maps which facilitate the study of evolutionary processes ranging from major structural changes such as genomic or intra chromosomal rearrangements to fine scale differences such as single nucleotide substitutions. Also the transfer of information like gene predictions directly from the closest sequenced reference species to the studied species is facilitated by comparative maps. Genetic markers are essential in constructing linkage maps facilitating comparative genomics. Thousands of genetic markers can also be applied in genome wide selection approaches, strategies that will substantially increase the rate of genetic gain in animal breeding.

By large scale identification of Single Nucleotide Polymorphisms (SNPs) and Structural Variants (SVs) we aimed to contribute to the development of a repository of genetic variants for farm animals. We targeted for the identification of sufficient SNPs allowing for construction of linkage maps, SNP genotyping based estimation of kinship or pedigree reconstruction and studies aiming for the detection of quantitative trait loci (QTL). For this purpose bioinformatics data pipelines were designed and validated to address the challenge of the cost effective identification of genetic markers in DNA sequencing data. Provided SNP detection strategies can be applied in almost any organism of interest without the requirement for a fully sequenced reference genome. By the lack of a reference genome SNPs are assigned to putative genomic locations by (comparative) mapping to closely related species. We also provide low cost strategies for obtaining sufficient sequencing data to reliably detect SNPs in species lacking a reference genome such as, at the time of these studies, turkey and duck. For the already sequenced chicken genome we show that limited paired end sequencing is sufficient to catch a glimpse of the abundance of structural variants.

High quality SNPs as a result of mining public sequencing data

In chapter 2 we identified SNPs in publicly available whole genome sequencing datasets consisting of output from traditional capillary sequencing platforms. We mined pig whole genome shotgun sequencing data by sequence alignment and clustering. Sequence clusters were assigned to genomic locations using publicly available BAC sequencing and BAC mapping data. We predicted thousands of SNPs within the sequence clusters and included a rough estimate of their genomic location. Genotyping of an animal panel proved that the overall performance of the SNPs identified by our genome shotgun sequence mining approach is comparable to those available in existing SNP repositories.

Cost effective approach of obtaining SNPs in unsequenced genomes

In chapter 3 we report on our first results using NGS, a technique enabling the quick and cost effective generation of a whole genome sequencing dataset. We show how this technique can be used to detect genetic variation in the, at that time, unsequenced turkey (Meleagris gallopavo) genome. To decrease sequencing costs we pooled the DNA of multiple individuals and sequenced a subset of the genome by creating a reduced representation library (RRL) of the pooled sample. By pooling the DNA of multiple individuals we reduced sequencing cost by handling only one (pooled) sample whereas the construction of a (RRL) resulted in a ten times reduction of the sequencing target while the thousands of obtained SNPs still cover, at intervals, the whole genome. Sequencing, assembly and SNP discovery were benchmarked by applying comparative genomics using the sequenced genome of the closely related species chicken. The quality of the SNPs and a correlation between genotype and sequence derived allele frequencies were determined by genotyping a selection of the identified SNPs.

First glimpse on the structural variation in the chicken genome

In chapter 4 we report on the application of paired end NGS for the detection of structural variation in four chicken breeds. Paired end sequencing is an extension on sequencing analysis that provides information about which pair of reads are coming from the outer ends of the same sequenced DNA fragment. We paired end sequenced reduced representation libraries (RRLs) of four chicken breeds that included genomic DNA fragments within the size range of 125-200bp. Paired end reads were optimally mapped to the chicken reference genome. SVs were identified as abnormally aligned read pairs that have an

orientation discordant from the reference genome or improper span sizes compared to the size-range of the RRL. We designed SV detection parameters to distinguish true structural variants from false positives due to RRL contaminants, sequencing and mapping errors or gaps in the chicken reference genome. Parameters were further optimized by validation of a small representative sample of SVs using PCR and traditional capillary sequencing. SVs were annotated and provide a first glimpse of the high resolution sequence map of chicken structural variation.

Genome wide SNP discovery and benchmarking

In chapter 5 we reimplemented our SNP detection approach described in chapter 3 according to the fast developments in NGS technology. We applied the reimplemented pipeline for obtaining SNPs to allow genotyping in wild duck (*Anas platyrhynchos)*. We show that the improvements in sequencing like higher sequence quality, paired end and longer read length facilitates variant discovery. A next-generation sequenced domesticated Peking duck reference genome consisting of tens of thousands of scaffolds and contigs served as a reference genome. Obtained SNPs were compared with two external SNP repositories that resulted from either SNP identification in the duck reference assembly or duck EST sequencing. A subset of high quality SNPs that were shared by our data and either of both external datasets was constructed, validated by genotyping and served to benchmark all SNPs that we identified within the wild ducks and those identified as sequence differences between wild duck and the domesticated Peking duck reference genome. Based on alignment results with the closely related chicken genome we estimate that we cover the complete duck genome with the SNPs identified.

What did we contribute and what can be expected in the near future

We developed SNP repositories which fulfill a requirement for SNPs to perform linkage analysis, comparative genomics QTL studies and ultimately GWA studies in a range of farm animals. We also set the first step in developing a repository for SVs in chicken, a relatively new genetic marker in animal sciences.

In chapter 6, I evaluate and discuss the results of our bioinformatics approaches for obtaining these high quality markers cost effectively. I discuss the exponentially growing amount of data in publicly available databases due to ongoing developments of new sequencing technologies which consequently result in less uniformity of the public data. Next I look back on our SNP

detection results and approaches using NGS technology. Furthermore I give my view on the significance of detecting structural variation in animal genomes and what methods can be considered. Finally I discuss what can be expected in the near future if emerging sequencing techniques and the development of efficient bioinformatics analysis solutions keep pace.

Samenvatting

Momenteel zijn voor landbouwhuisdieren de aantallen beschikbare genetische merkers niet toereikend en voor veel diersoorten zijn ze zelfs helemaal niet voorhanden. Genetische merkers zijn echter essentieel voor het ontwikkelen van een methode om genetische factoren te identificeren die te maken hebben met gevoeligheid voor ziekte, resistentie tegen ziekte en het algeheel welbevinden en de prestaties van landbouwhuisdieren. Door middel van het vergelijken van genomen binnen een soort of van aan elkaar verwante soorten kunnen veranderingen in het genoom worden onthuld. Samen met de registratie van kenmerken over meerdere generaties kan met deze kennis de genetische achtergrond van economisch en biologisch belangrijke eigenschappen worden geïdentificeerd. Het beschikbaar zijn van genetische kaarten en de genoom sequentie, welke in zekere mate geconserveerd is tussen verwante soorten, stelt ons in staat genomen te vergelijken en verschillen gedetailleerd in kaart te brengen. De veranderingen door toedoen van genoom evolutie variëren van kleinschalige veranderingen zoals de substituties van enkele basen tot intra chromosomale of genomische herschikking. In kaart gebrachte genoom vergelijkingen worden gebruikt om kennis als bijvoorbeeld genvoorspellingen van de soort waarvan de genoom sequentie bekend is, toe te passen op de soort in studie.

Een genetische merker is een variatie in het erfelijk materiaal (DNA) waarvan de overerving te volgen is. Genetische merkers zijn onmisbaar bij het in kaart brengen van de verbondenheid van genomische regio's tot elkaar (linkage studies), wat belangrijke informatie is bij het vergelijken van genomen. Merkers worden ook gebruikt in het volgen van erfelijke eigenschappen in een stamboom. Tegenwoordig worden in de fokkerij duizenden genetische merkers toegepast om bij selectieprocedures zoveel mogelijk genomische informatie in de afweging mee te nemen. Met deze aanpak kan in de verbetering van de genetische samenstelling van landbouwhuisdieren per generatie-interval meer vooruitgang worden geboekt.

Met de grootschalige identificatie van base substituties (SNPs) en structurele varianten (SVs) in landbouwhuisdieren hebben we willen bijdragen tot de ontwikkeling van een verzameling genetische variatie, welke als merkers gebruikt kunnen worden. We stelden ons tot doel voldoende SNPs in te identificeren welke het mogelijk maken genetische kaarten te construeren, stambomen van een populatie te reconstrueren en de verwantschap tussen dieren te kunnen bepalen. Ook wilden we studies faciliteren waarin locaties op

het DNA worden geïdentificeerd welke een effect hebben op een bepaalde erfelijke eigenschap van een dier (QTL studies). Om ons doel te kunnen bereiken zijn, door toepassing van bio-informatica, datapijplijnen ontworpen en gevalideerd. Hiermee zijn we de uitdaging aangegaan de genetische variatie in publiek beschikbare DNA sequentie analyse data zo kosteneffectief mogelijk identificeren. Onze aanpak van SNP detectie kan worden toegepast in nagenoeg elk organisme zonder dat daarvoor de volledige DNA sequentie van een referentiegenoom beschikbaar moet zijn. Het ontbreken van een referentiegenoom wordt gecompenseerd door SNPs aan genomische locaties toe te wijzen met behulp van een beschikbare genoom vergelijking met een verwante soort, waarvoor wel genomische informatie beschikbaar is. Daarnaast laten we strategieën zien waarmee tegen lage kosten voldoende DNA sequentie data kan worden verkregen om SNPs te detecteren in soorten zonder referentiegenoom en onvoldoende DNA sequentie data, zoals ten tijde van onze studie kalkoen en eend. Tenslotte laten we zien dat een beperkte sequentie analyse op uiteinden van fragmenten van het kippen genoom voldoende is om een eerste indruk te krijgen van de aanwezigheid van structurele variatie.

Kwalitatief goede SNPs als resultaat van het doorspitten van publieke sequencing data.

In Hoofdstuk 2 hebben we SNPs geïdentificeerd in publieke genoom sequencing data afkomstig van traditionele capillaire sequentie analyse platformen. We hebben quasi-random sequentie-analyse-data van varken geanalyseerd door overeenkomstige DNA sequenties te zoeken en deze te clusteren. Deze DNA sequentie clusters werden toegekend aan genomische locaties indien ze homoloog waren aan additionele publieke sequentie data waarvan de genomische locatie bekend was. We voorspelden duizenden SNPs in de DNA sequentie clusters en konden aan de verkregen SNPs een grove indicatie van de genomische positie toekennen. Met het genotyperen van een selectie dieren hebben we bewezen dat de prestaties van de SNPs die we met onze aanpak voorspelden vergelijkbaar is met reeds beschikbare SNPs.

Kosten effectieve aanpak voor het verkrijgen van SNPs in genomen waarvoor geen sequentie data beschikbaar is.

In hoofdstuk 3 rapporteren we over onze eerste resultaten met een nieuwe generatie sequentie-analyse-apparatuur (NGS), waarmee snel en kosteneffectief een sequentie dataset van een heel genoom kan worden gegenereerd. We laten

zien hoe deze techniek kan worden gebruikt voor detectie van SNPs in het kalkoen genoom, waarvoor in die tijd nog geen genoom sequentie beschikbaar was. Omwille van lage sequencing kosten hebben we het DNA van meerdere kalkoenen samengevoegd en van fragmenten verspreid over het genoom de sequentie bepaald. Het samenvoegen beperkte de sequentie analyse tot één DNA sample. De complexiteit van dit sample hebben we met een tienvoud gereduceerd door geen sequentie analyse op het hele genoom uit te voeren, maar verspreid over het genoom kleine fragmenten. Hiermee vergrootten we de kans dat we van meerdere kalkoenen DNA-sequentie van dezelfde genomische locatie hebben. De vergelijkingen van de sequenties afkomstig van dezelfde genomische locaties leverde duizenden kalkoen SNPs op. Het kippengenoom, waarvan de DNA sequentie al bekend is, hebben we gebruikt om de sequentie analyse, het reconstrueren van de genoom fragmenten en de detectie van SNPs te staven. De geïdentificeerde SNPs bleken in intervallen het kippengenoom geheel te bedekken. De kwaliteit van de SNPs en of de frequentie waarin de variatie gevonden is in de sequentie data overeenkomt met die in een populatie kalkoenen, is vastgesteld door met een selectie van de geïdentificeerde SNPs een groep kalkoenen te genotyperen.

Een eerste indruk van de structurele variatie in het kippen genoom

In hoofdstuk 4 beschrijven we de toepassing van paired-end NGS voor het detecteren van structurele variatie in vier kippen rassen. Paired-end sequentie analyse geeft informatie over welk paar sequenties in sequentie data afkomstig is van de uiteinden van hetzelfde op sequentie geanalyseerde DNA fragment. We hebben verspreid over het genoom van vier kippenrassen paired-end sequentie analyse gedaan op fragmenten tussen de 125 en 300 basenparen in lengte. De optimale locaties van beide sequenties van ieder sequentie-paar werden bepaald op het kippen referentiegenoom. Structurele varianten (SVs) werden herkenbaar doordat sequenties van een sequentie paar hun ideale positie open het referentiegenoom vonden in een grotere of kleinere afstand van elkaar dan de verwachtte afstand (125-300 bp). SV detectie parameters werden opgesteld om echte SVs van vals positieven (veroorzaakt door fragment contaminatie, fouten in sequentie en locatie bepaling of fouten in het referentiegenoom) te onderscheiden. De parameters werden verder geoptimaliseerd door de validatie van een kleine representatieve steekproef met behulp van PCR en traditioneel capillaire sequentie analyse. SVs werden vervolgens geannoteerd en geven de eerste resultaten weer van het in kaart

brengen van structurele variatie in het kippen genoom.

Genoom wijde SNP detectie en prestatie vergelijkingen

In hoofdstuk 5 hebben we de SNP detectie methode, zoals beschreven in hoofdstuk 3, aangepast op de snelle ontwikkelingen in NGS technologie. We hebben de aangepaste methode toegepast in eend voor het verkrijgen van SNPs welke het genotyperen van wilde eenden mogelijk maken. We laten zien dat de ontwikkelingen in sequentie analyse technologie, zoals hogere kwaliteit van sequenties, gepaarde uiteinden en langere sequenties, de opsporing van variatie ten goede komen. Het genoom van de gedomesticeerde eend hebben we als referentiegenoom gebruikt. Dit genoom is met NGS technologie geanalyseerd en bestond uit duizenden langere en korte sequentie fragmenten. Verkregen SNPs werden vergeleken met SNPs uit twee externe SNP verzamelingen welke het resultaat waren van SNP identificatie bij het maken van het referentiegenoom van de eend ofwel de sequentie analyse op het deel van het eendengenoom dat tot expressie komt (EST-sequencing). De verzameling SNPs die in onze resultaten voorkwamen en ook in één van de andere twee verzamelingen, werden gevalideerd middels genotyperen. Deze verzameling gemeenschappelijke SNPs vormde de maatstaf waaraan we onze volledige SNP resultaten gemeten hebben. De volledige verzameling bestaat uit SNPs die we geïdentificeerd hebben als verschillen tussen wilde eenden en SNPs die we identificeerden als verschillen tussen wilde eend en het genoom van de gedomesticeerde eend. Resultaten van het positioneren van eenden SNP op het kippen genoom suggereren dat we SNPs hebben gevonden verspreid over het hele eenden genoom.

Wat hebben we bijgedragen en wat kan worden verwacht in de nabije toekomst

We hebben SNP verzamelingen ontwikkeld welke voorzien in een vraag naar SNPs om linkage studies, genoom vergelijkingen en QTL-studies te doen in landbouwhuisdieren. Als onze SNP verzamelingen verder worden aangevuld, dan zijn studies mogelijk waarin honderdduizenden genetische variaties worden getest op correlatie met een overerfbare eigenschap (GWA studie). Verder hebben we de eerste stap gezet in het ontwikkelen van een verzameling SVs in kip, een relatief nieuwe genetische merker in de dierwetenschappen. In hoofdstuk 6 evalueer en bediscussieer ik de resultaten van onze bio-informatica aanpak om kosteneffectief aan kwalitatief goede merkers te komen. Ik ga in op de exponentieel groeiende hoeveelheid sequentie data in publieke databases

door de voortschrijdende ontwikkelingen van nieuwe sequentie analyse technieken welke als gevolg hebben dat data minder uniform is geworden. Ook kijk ik terug op onze SNP detectie resultaten en onze aanpak waarbij gebruik werd gemaakt van NGS. Vervolgens geef ik mijn mening over het belang van de detectie van structurele variatie in dierlijke genomen en welke methoden er overwogen kunnen worden. Tenslotte bediscussieer ik wat in de nabije toekomst kan worden verwacht als de ontwikkeling van efficiënte bio-informatica analyses en het tempo waarin sequentie analyse technologieën blijven opduiken gelijke tred houden.

Dankwoord

Begin 2004 kreeg ik de kans, op detacherings basis, voor een periode van twee keer twee jaar als bio-informaticus te gaan werken bij de leerstoelgroep fokkerij en genetica. Daarvoor had ik in de rol van analist en bio-informaticus al ervaring opgedaan in het hanteren en analyseren van micro-array resultaten. Toenmalige collega's zagen mijn vertrek naar Wageningen als een definitief afscheid, zelf had ik toen nog de intentie terug te keren. Wat me aantrok tot de vakgroep in Wageningen was mijn vermoeden dat de bio-informatica daar al verder was dan in Lelystad en omdat er mogelijkheden waren een promotie-traject in te gaan. Ondanks een moeizame start waarin een mogelijk hoofdstuk van mijn proefschrift verloren ging en ik vastgesteld had dat de bio-informatica veel minder ver was dan ik vermoedde braken de betere jaren van het promotie-traject aan. Er kwam sequentie data...eerst uit het publieke domein maar daarna kregen we ook toegang tot de nieuwste generatie sequentie analyse apparatuur. Bovendien werd wat geld door de groepen in Lelystad en Wageningen bijeen gelegd zodat ik naar eigen inzicht een computercluster kon bouwen waarbij ik Pieter Kroon (proceskunde) nogmaals wil bedanken voor het regelen van een locatie. Nu aan de voorwaarden “data” en “rekenkracht” was voldaan, konden we los. Met inzet van Arun Kommadath (toen masters student nu AIO in Lelystad) en Marisol del Rosario (toen stagiaire nu collega aan de andere kant van de loopbrug) hebben we de identificatie van sequentie variatie in publieke varkens genoom sequencing data volledig weten te automatiseren. Helaas heeft de pijplijn na het verschijnen van de publicatie niet meer kunnen draaien door toedoen van de stroom Next Generation Sequencing (NGS) data maar toch bedankt voor jullie inzet. Dat ik het voorrecht had als één van de eersten in Nederland met NGS data te werken, kwam door een samenwerking tussen Martien Groenen en Johan de Dunnen waar ik beiden voor wil bedanken. De aard, omvang en complexiteit van de data was aanleiding kennis te maken met de Huygens-supercomputer in Nederland. NBIC deed me het idee van het gebruik van de supercomputer faciliteiten aan de hand en NCF en SARA maakten het financieel en praktisch mogelijk.
Maar met alleen wat stukken hardware heb je nog geen op maat gesneden data analyse pijplijn in handen. Daarvoor is een besturingssysteem nodig die aan de ene kant de hardware efficiënt aanstuurt en aan de andere kant de gebruiker bij voorkeur alle vrijheid geeft software te ontwikkelen en geen geheimen kent. Linux voldoet aan deze criteria waarvoor ik de ontwikkelaars oneindig dankbaar ben, want zonder dit besturingssysteem had ik dit werk niet willen en

kunnen doen. Het idee om met linux aan de slag te gaan, inclusief installatie CDs en een beschrijving over hoe je een netwerk opzet kreeg ik in 1998 van Victor Huisman. Victor, wellicht dat je toen geen vermoeden had van waar je me toe hebt aangezet, maar het idee op zich was goed. En mocht je nog vragen hebben over linux...
En dan nu weer even terug naar Zodiac want naast die 1152 pixels horizontaal en 864 pixels verticaal was er op Zodiac natuurlijk ook nog de groep. Schoof ik mijn stoel even opzij dan zag ik bijvoorbeeld de achterkant van 1024x768 pixels van het scherm van mijn kamergenoot Haisheng Nie. Ondanks dat we vrijwel dagelijks in Area 51 (room with the deadly force) zaten, zagen we elkaar door de opstelling van de monitoren weinig. Tot interessante en bij vlagen humoristische conversaties kwam het gelukkig zeer regelmatig. Wel moet ik me verontschuldigen voor de storende aanloop die ik had van collega's met vragen over hun en vooral mijn scripts.
Ook was er een keer een dag waarop we als groep leerden hoe zinloos geweld en hoe belangrijk een stapel dozen kan zijn. Voor diegenen die mij nog steeds als één van de mede-schuldigen van het zinloze geweld zien:"SORRY"het waren Robert,Pieter en Koen. Verder moet ik de beleggers in "The Red Pool" helaas teleurstellen; het fonds bestaat niet meer maar bedankt voor jullie kennis (Jan) en vertrouwen.
Collega's, met name die uit de green-room, bedankt voor de luchtige (zo luchtig als grond is wanneer je er met een tuinklauw goud doorheen bent geweest), verlichtende en informatieve gesprekken tijdens de pauzes. Het informatieve zat hem in de aard van de gesprekken, waaraan kon worden afgeleid welke dag van de week het was. Wat me tegenvalt is dat jullie het hot-swappen zijn verleerd en zelf praktiseer ik deze koffie-tap-techniek ook al maanden niet meer. Zou toch jammer zijn als dit verloren ging. Verder wil ik Richard, Bert, Tineke en Sylvia bedanken voor de geleverde bijgedragen aan de in-vitro validatie van de in-silico voorspelde variatie. Het deed me goed wanneer jullie weer met resultaten kwamen die bepaald niet ruk waren. Hendrik-Jan en Nikkie wil ik bij deze respectievelijk bedanken voor de leuke discussies en de kritische blik op mijn werk. Het lijkt erop dat jullie wèl begrepen hebben wat de pinguïn in een computer voor je kan doen, maar om als moeder je kinderen naar linux distributies te gaan noemen....
Over mijn supervisie had ik niet te klagen. Mijn dageweekelijkse begeleider Martien is een ras-echt optimist en zij coaching kon ik wel gebruiken als ik mijn resultaten weer eens in een wel heel negatief licht had geplaatst. Daarnaast wil ik mijn begeleider op afstand (Mari) bedanken voor zijn visie en

commentaar op mijn hoofdstukken. Ik heb het altijd als prettig ervaren met je in discussie te gaan omdat je het geheel goed overzag en vanuit een meer biologisch perspectief.
Ondanks dat ik nogal eens een andere kijk op zaken had, heb ik de leerstoelgroep fokkerij en genetica als een goede voedingsbodem ervaren om me op te ontwikkelen. Naast een gezonde druk om te presteren hing er altijd de ontspannen sfeer van een kerstmarkt. Van het sociale programma buiten de werktijden heb ik helaas niet veel meegekregen omdat de zorgtaken in de avonduren thuis de afgelopen vier jaar er niet om logen. Met een dochter voor wie de op fysiek gebied de meest vanzelfsprekende ontwikkelingen niet vanzelfsprekend waren, zijn mijn echtgenoot en ik jarenlang door diepe dalen gegaan. Gelukkig hebben we met hulp van de zorgverleners van Groot Klimmendaal (waarvoor heel veel dank) sinds afgelopen voorjaar een gezinsleven waarin niet de hele dag in het teken staat van eetproblemen. Ondanks de moeilijke vier jaren is het toch gelukt tijd aan onderzoek te besteden en voldoende resultaten te behalen voor het samenstellen van dit proefschrift. Dit was niet gelukt zonder de grenzeloze steun van mijn echtgenote, Marleen. De pedagogische zorgvuldigheid en het geduld waarmee jij met onze kinderen, Indra en Jorim, omgaat heeft de sfeer binnen het gezin ,ondanks de problemen, altijd goed gehouden. De goede sfeer heeft ervoor gezorgd dat ik me in alle uren die ik had, me volledig op dit werk kon concentreren.
Het belang van het creëren van de goede omgeving om te kunnen presteren was me al vanuit mijn opvoeding bekend. Pa en ma wat hebben jullie beiden goed aangevoeld wat Cor, Lyanda en ik nodig hadden in onze school- puber- en studententijd om goed voorbereid, zelfstandig en succesvol verder te kunnen. Ook heb ik het altijd bijzonder gevonden dat Cor Lyanda en ik, zelfs als pubers, elkaar een rustige omgeving gunden om te studeren, waarvoor ik broer en zus bij deze wil bedanken.
In de avonduren van het afgelopen half jaar heb ik jou (Marleen) wellicht iets te veel met rust gelaten. Na een vrolijke begroeting bij thuiskomst van jou en de kinderen waren alleen het eten en het de kinderen in bad en naar bed brengen onze gezamenlijke activiteiten. Dat moesten we de komende tijd maar weer eens goed maken.

Hinri

About the Author

Hindrik Harm Dirk Kerstens was born on February 5th 1975 in Groningen,The Netherlands. He grew up in Haren and finished his HAVO at the the Zernike College in Haren in 1993. In 1997 he finished his study Biotechnology at the Hanzehogeschool van Groningen. During his study, his specialization was molecular biology and he did his intern ship "Transposon tagging in *Arabidopsis*" at the Centre for Plant Breeding and Reproduction Research (CPRO-DLO) in Wageningen. After his study he immediately started working as a lab technician at the high containment labs of the Institute of Animal Science and Health (ID-DLO) in Lelystad. The late hours in Lelystad he spent in developing linux computing skills driven by the frustrating limitations of another operating system he was using at that time and which many of you still are using. Acquired knowledge he applied in the lab of doctor Rob Moormann by automating data analyses and computer-aided drafting and design of cloning strategies. In 2002 he moved to the micro-array lab of the Animal Sciences Group to continue working with Marcel Hulst but this time under supervision of doctor Mari Smits. In this lab he autodidacticly aquired knowledge about handling and analysing large datasets and building linux computer clusters. On the first of April 2005 he went to the chair group Animal Breeding and Genetics at the Wageningen University, where he worked as bio-informatician under supervision of professor Martien Groenen. In addition to supporting the lab by creating and maintaining data analyses pipelines and a linux computer cluster facility, he performed the scientific work described in this thesis. Since February 15th 2010, he is working as bioinformatician in the molecular biology group of professor Henk Stunnenberg at the Radboud University of Nijmegen.

List of Publications

Papers in refereed journal

van Bers NEM, van Oers K, Kerstens HHD, Dibbits BW, Crooijmans RPMA, Visser ME, Groenen MAM: **Genome-wide SNP detection in the great tit Parus major using high throughput sequencing.** Mol Ecol 2010, **19** Suppl 1:89-99.

Niewold TA, van der Meulen J, Kerstens HHD, Smits MA, Hulst MM: **Transcriptomics of enterotoxigenic Escherichia coli infection. Individual variation in intestinal gene expression correlates with intestinal function**. *Vet Microbiol* 2010, **141**:110-114.

Megens H, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH, Hanotte O, Groenen MAM: **Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken**. *BMC Genet* 2009, **10**:86.

Kerstens HHD, Crooijmans RPMA, Veenendaal A, Dibbits BW, Chin-A-Woeng TFC, den Dunnen JT, Groenen MAM: **Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey**. *BMC Genomics* 2009, **10**:479.

Amaral AJ, Megens H, Kerstens HHD, Heuven HCM, Dibbits B, Crooijmans RPMA, den Dunnen JT, Groenen MAM: **Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome**. *BMC Genomics* 2009, **10**:374.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z, Kerstens HH, Law AS, Megens H, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology**. *PLoS ONE* 2009, **4**:e6524.

Kerstens HH, Kollers S, Kommadath A, Del Rosario M, Dibbits B, Kinders SM, Crooijmans RP, Groenen MA: **Mining for single nucleotide polymorphisms in pig genome sequence data**. *BMC Genomics* 2009, **10**:4.

Hulst M, Kerstens H, de Wit A, Smits M, van der Meulen J, Niewold T: **Early transcriptional response in the jejunum of germ-free piglets after oral infection with virulent rotavirus**. *Arch Virol* 2008, **153**:1311-1322.

Gut-Winiarska M, Jacobs L, Kerstens H, Bienkowska-Szewczyk K: **A highly specific and sensitive sandwich blocking ELISA based on baculovirus expressed pseudorabies virus glycoprotein B**. *J Virol Methods* 2000, **88**:63-71.

Papers submitted or in preparation

Kerstens HHD, Kraus RHS, van Hooft P, Crooijmans RPMA, van der Poel J, Elmberg J, Vignal A, Huang Y, Li N, Prins HHT, Groenen MAM: **Genome wide SNP discovery and analysis in mallard (Anas platyrhynchos).** In preparation.

Kerstens HHD, Crooijmans RPMA, Dibbits BW, Vereijken A, Okimoto R, Groenen MAM:**Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries.** Submitted.

Conference abstracts

Crooijmans RPMA, Kerstens HHD, Amaral AJ, Veenendaal A, Dibbits BW, Chin-A-Woeng TFC, den Dunnen JH, Groenen MAM: **Identification Of Porcine And Turkey SNPs By High Parallel Sequencing On A Solexa Sequencing Platform.** *In: Plant and Animal Genome XVI, 12-16 January 2008, San Diego, USA.*

Groenen MAM, Kerstens HHD, Amaral AJ, Veenendaal A, Chin-A-Woeng TFC, Dunnen J, Crooijmans RPMA: **LD In Pigs: Differences Between China And Europe.** *In: Plant and Animal Genome XVI Conference, 12-16 January 2008, San Diego, USA.*

Megens HJWC, Crooijmans RPMA, Harlizius B, Kollers S, Mullaart E, Jungerius BJ, Kerstens HHD, Groenen MAM: **A Perl Based Data Mining Workflow For Animal Breeding - From Phenotype To SNP.** *In: Plant and Animal Genome XVI Conference, 12-16 January 2008, San Diego, USA.*

Amaral AJ, Kerstens HHD, Megens HJWC, Dibbits BW: **Genome wide SNP discovery in pig using 1G Genome Analyzer: How to play around with sequence length, quality level and mapping qualities.** *In: Conference of the*

31st International Society for Animal Genetics, 20-24 July 2008 Amsterdam, The Netherlands.

Ramos A, Amaral AJ, Kerstens HHD, Bendixen C, Hedegaard J, Rohrer G, Smith T, Tassell C, Taylor JF: **High throughput SNP discovery and validation in the Pig: towards the development of a high density swine SNP chip.** *In: Book of abstracts of the XXXI Conference of the International Society for Animal Genetics, 20 - 24 July 2008, Amsterdam. The Netherlands.*

Presentations

Kerstens HHD, Dibbits BW, Crooijmans RPMA, Groenen MAM: **Paired-End Next-generation DNA sequencing of Reduced Representation Libraries for the analysis of structural variation in the Chicken genome.** In *Genomics for Animal Health: Outlook for the Future-Bioinformatics and Data Handling for Next Generation Sequencing, 12th October 2009, Paris, France*

Kerstens HHD, Crooijmans RPMA, Groenen MAM: **Paired-End Next-generation DNA sequencing of reduced representation libraries for the analysis of structural variation in the Chicken genome.** *In: Conference of the 2nd Belgium-Netherlands-Luxembourg Next-generation sequencing Users Meeting, 7th July 2009, Utrecht, The Netherlands.*

Kerstens HHD, Amaral AJ, Megens HJWC, van Bers NEM, Kraus HS, Crooijmans RPMA, den Dunnen J, Groenen MAM: **High throughput SNP discovery in unsequenced genomes using 2nd generation sequencing technology.** *In: Benelux Bioinformatics Conference 2008, December 14-15 2008, Maastricht, The Netherlands.*

Posters

Kerstens HHD, Amaral AJ, Megens HJWC, Hemmatian K, Crooijmans RPMA, den Dunnen J, Groenen MAM: **High throughput SNP discovery in unsequenced genomes using 2nd generation sequencing technology.** *In: EADGENE 4th annual meeting on Animal Genomics Research, 9-12 June 2008, Edinburg, UK.*

Kerstens HHD, Groenen MAM: **Analyzing orthologue predictions using conserved synteny.** *In: 2nd EADGENE days, 29-30 June 2006,Oslo, Norway.*

Training and Supervision Plan

The Basic Package (3 credits)

WIAS Introduction Course	2008
Course on philosophy of science and/or ethics	2009

Scientific Exposure (13 credits)

International conferences

Eadgene days, Oslo (NO), May 29-30	2006
Eadgene days, Utrecht (NL), June 4-8	2007
Bioinformatics OpenSource Conference, Vienna (AT), July 19-20	2007
Eadgene days, Edinburgh (UK), June 9-12	2008
Benlux Bioinformatics Conference, Maastricht (NL), December 15-16	2008
Eadgene days, Paris (FR), October 13-15	2009
Second Next-Generation Sequencing Users Meeting, Utrecht (NL), July 7	2009

Seminars and workshops

NBIC-ISNB, Amsterdam (NL), April 16-19	2007
WIAS science day, Wageningen (NL), March	2008
WIAS science day, Wageningen (NL), March 12	2009
Bioinformatics and Data Handling for Next Generation Sequencing, Paris (FR), October 12	2009

Presentations

Orthologue predictions using conserved synteny, poster presentation, Eadgene days, Oslo (NO), 29-30 June	2006
SNP discovery in unsequenced genomes by NGS, poster presentation, Eadgene days, Edinburgh (UK), 9-12 June	2008
High throughput SNP discovery using NGS, oral presentation, BBC2008, Maastricht (NL), 15-16 December	2008
SNP discovery in unsequenced genomes using NGS, oral presentation, WIAS Sc. Day, Wageningen (NL), 12 Mar	2009
PE-NGS of RRL for the analysis of SVs in chicken, oral presentation, NGS users meeting, Utrecht (NL), 7 July	2009
PE-NGS of chicken RRLs for the detection of SVs,oral presentation, Eadgene NGS, Paris (FR), 12 October	2009

In-Depth Studies (7 credits)

Disciplinary and interdisciplinary courses

C programming, AT-computing, Utrecht (NL) April 10-15	2008
Ensembl Workshop for Developers, Cambridge University, Cambridge (UK) April 29-May 1	2009
C++ programming ,AT-computing, Nijmegen (NL) May 28-June 3	2009
Visualisation, integration and biological interpretation of -omics data, VLAG, Wageningen (NL) October 26-28	2009

Advanced statistics courses

AIO course Biostatistics for Researchers, Julius Center, Utrecht (NL) October 27-November 7	2008

Professional Skills Support Courses (4 credits)

Introduction to R for Statistical Analysis, WIAS, Wageningen (NL) April 21-22	2008
Course Techniques for Scientific Writing July 1-4	2008
Effective behaviour in your professional surroundings February 23 + March 23	2009
Project- and Time Management	2009

Research Skills Training (1 credit)

Design,build and maintenance of 22 core shared computation facility (2006-2009)	2009

Didactic Skills Training (3 credits)

Supervising students	2008

Supervising theses

Supervising HBO Stagiaire	2007
Supervising minor thesis	2006

Management Skills Training (1 credit)

Membership (advisor) Facility Committee	2009

Total credits: 32 *(one credit equals a study load of approximately 28 hours)*

Colophon

The research as described in this thesis was funded by European Union grant FOOD-CT-2004-506416 (Eadgene).

This thesis was printed by Wöhrmann Print Service, Zutphen, the Netherlands