

Università degli Studi della Tuscia
Dipartimento di Scienze Ecologiche e Biologiche

Dottorato in Genetica e Biologia Cellulare

XXIV Ciclo

**Second Generation Sequencing tools: from
de novo genome assembly to quantitative
genetics studied by dense SNP chips**

Dottorando:	Giordano Mancini
Relatore:	Dr. Lorraine Pariset
Coordinatore:	Prof. Giorgio Pranterà

*Non sono fini, termini, margini, muraglia che ne defrodino e suttragano la infinita copia de le cose.
Indi feconda è la terra et il suo mare; indi perpetuo è il vampo del sole: sumministrandosi eternamente esca a
gli voraci fuochi et humori á gl'attenuati mari: perché dall'infinito sempre nova copia di materia sottonasce*

Giordano Bruno

Table of Contents

1	FOREWORD.....	1
2	THEORETICAL BACKGROUND.....	2
2.1	INTRODUCTION.....	2
2.1.1	<i>Sanger Sequencing</i>	4
2.1.2	<i>Second Generation Sequencing</i>	5
2.1.2.1	Roche 454.....	6
2.1.2.2	Illumina.....	7
2.1.2.3	SOLiD.....	8
2.1.2.4	Third Generation Sequencing.....	9
2.1.3	<i>De Novo assembly algorithms</i>	9
2.1.3.1	Introduction	9
2.1.3.2	The Overlap-Layout-Consensus approach.....	10
2.1.3.3	The De Bruijn graph approach.....	11
2.1.3.4	The MSR-CA assembler.....	13
2.2	SNP PANELS.....	15
2.2.1	<i>Introduction</i>	15
2.2.2	<i>The Illumina BovineSNP50 chip</i>	16
2.3	STATISTICAL METHODS.....	17
2.3.1	<i>Infinitesimal model and Quantitative Trait Locus model (QTL)</i>	17
2.3.2	<i>Linkage Disequilibrium</i>	20
2.3.3	<i>Population genetics</i>	23
2.3.3.1	Introduction.....	23
2.3.3.2	Population structure.....	24
2.3.3.3	F statistics.....	27
2.3.4	<i>Genome-wide scan (GWAS) using SNP panels</i>	28
2.3.4.1	Genotyping errors.....	29
2.3.4.2	Estimation of kinship matrix from dense SNP data.....	30
2.3.4.3	Analysis of structured populations.....	31
2.3.4.4	Genomic control.....	32
2.3.4.5	Analysis of family data.....	33
2.3.4.6	The FASTA method.....	34
3	DE NOVO ASSEMBLY OF WATER BUFFALO GENOME.....	36
3.1	INTRODUCTION.....	36
3.2	THE WATER BUFFALO DATA SET.....	36
3.3	ROCHE 454 PREPROCESSING.....	38
3.4	ERROR CORRECTION AND SUPER-READS CREATION.....	39
3.4.1	<i>Illumina Mate Pair preprocessing</i>	40
3.4.2	<i>De Bruijn graph construction and error correction</i>	42
3.4.3	<i>Jump libraries preprocessing after error-correction</i>	43
3.4.4	<i>Super-reads creation</i>	44
3.5	GENOME ASSEMBLY USING SUPER-READS.....	46
3.6	GUIDED ASSEMBLY ALIGNING ON BOS TAURUS GENOME.....	47
3.7	COMPUTATIONAL COST.....	48

4	CATTLE QUANTITATIVE GENETICS.....	50
4.1	THE SELMOL DATA SAMPLE.....	50
4.1.1	<i>Animals and phenotype data</i>	50
4.1.2	<i>SNP panel and genotyping</i>	50
4.2	CANDIDATE GENES IN ITALIAN BROWN BREED.....	50
4.2.1	<i>Introduction</i>	50
4.2.2	<i>Candidate gene selection</i>	51
4.2.3	<i>Results</i>	52
4.2.3.1	Data Analysis.....	52
4.2.3.2	SNP association to milk yield.....	53
4.2.3.3	SNP association to fat yield.....	55
4.2.3.4	SNP association to protein yield.....	56
4.2.3.5	SNP Haplotypes	56
4.2.4	<i>Discussion</i>	57
4.3	ASSESSMENT OF POPULATION STRUCTURE IN FIVE ITALIAN CATTLE BREEDS.....	58
4.3.1	<i>Introduction</i>	58
4.3.2	<i>Results</i>	59
4.3.2.1	Preprocessing.....	59
4.3.2.2	Data analysis.....	60
4.3.2.3	Genomic kinship clusters.....	61
4.3.2.4	Bayesian inference	64
4.3.2.5	Fixation Index (Fst).....	66
4.3.3	<i>Discussion</i>	68
4.4	IDENTIFICATION OF A SHORT REGION IN CHROMOSOME 6 AFFECTING CALVING TRAITS IN PIEMONTESE CATTLE.....	70
4.4.1	<i>Introduction</i>	70
4.4.2	<i>Results</i>	71
4.4.2.1	Preprocessing.....	71
4.4.2.2	Genome-wide scan.....	72
4.4.2.3	SNP discovery, genotyping and analysis.....	75
4.4.3	<i>Discussion</i>	77
5	SUMMARY OF RESULTS	80
5.1	WATER BUFFALO GENOME ASSEMBLY.....	80
5.2	CANDIDATE GENES IN ITALIAN BROWN.....	80
5.3	ASSESSMENT OF POPULATION STRUCTURE IN FIVE ITALIAN CATTLE BREEDS	81
5.4	IDENTIFICATION OF A SHORT REGION IN CHROMOSOME 6 AFFECTING CALVING TRAITS IN PIEMONTESE CATTLE.....	81
5.5	CONCLUSIONS.....	82
6	REFERENCES.....	I
	ACKNOWLEDGEMENTS.....	XII

List of Figures

Figure 1: Genomic DNA is broken into a collection of fragment.....	3
Figure 2: The ends of each fragment are sequenced.....	3
Figure 3: The sequence reads are assembled together based on sequence similarity.....	3
Figure 4: Sanger sequencing workflow.....	5
Figure 5: Second generation sequencing read types.....	6
Figure 6: 454 sequencing workflow.....	7
Figure 7: Illumina sequencing workflow.....	8
Figure 8: Elements of a DNA assembly.....	10
Figure 9: Overlap graph for a bacterial genome.....	11
Figure 10: Schematic diagram of the de Bruijn graph implementation.....	12
Figure 11: MSR-CA workflow.....	13
Figure 12: An example of a read whose super read has two k-unitigs.....	14
Figure 13: Single Nucleotide Polymorphisms.....	15
Figure 14: Illumina SNP50 Beadchip.....	16
Figure 15: Distribution of additive (QTL) effects from pig experiments.....	17
Figure 16: Linkage mapping.....	19
Figure 17: Pairwise Linkage Disequilibrium.....	21
Figure 18: Illustration of LD blocks and associated tag SNPs.....	23
Figure 19: Structured populations.....	31
Figure 20: fascinating Olympia.....	37
Figure 21: 454 preprocessing workflow.....	38
Figure 22: Roche 454 preprocessing results.....	38
Figure 23: Illumina jump libraries preprocessing workflow.....	39
Figure 24: Origin and Alignment of Inward (innies) and Outward Facing (outties) reads in Mate Pair libraries.....	40
Figure 25: Alignment of Illumina jump libraries on Bos taurus genome.....	41
Figure 26: Alternative jump libraries preprocessing workflow.....	42
Figure 27: Results of mapping on Bos taurus after error correction- 1.....	42
Figure 28: Results of mapping on Bos taurus after error correction - 2.....	43
Figure 29: Super-reads creation workflow.....	43
Figure 30: Super-reads creation results.....	44
Figure 31: Super-reads dimension.....	44
Figure 32: CABOG assembly workflow.....	45
Figure 33: Guided assembly results.....	46
Figure 34: Breed-wise genetic kinship coefficients.....	60
Figure 35: Classical Multidimensional Scaling (MDS) plot of genomic distance for all five breeds.....	61
Figure 36: MDS plot of genomic distance for dairy breeds.....	62
Figure 37: MDS plot of genomic distance for beef breeds.....	63
Figure 38: Summary plot of Q estimates for K=2,3,4,5.....	64
Figure 39: Genome wide Manhattan plots of Fixation index (Fst).....	66
Figure 40: Fst calculated for all markers on BTA6.....	67
Figure 41: Histogram of EBV values for direct calving ease	70

Figure 42: Manhattan plot of Log significance levels of SNPs associated with direct calving ease.....	72
Figure 43: Genes in the BTA6 peak.....	72
Figure 44: Pairwise LD between SNPs on BTA6 peak.....	74

List of Tables

Table 1: Italian Brown candidate genes.....	49
Table 2: Italian Brown SNPs statistics.....	51
Table 3: SNPs associated to milk yield.....	52
Table 4: SNPs associated to fat yield.....	52
Table 5: SNPs associated to protein yield.....	53
Table 6: SNP haplotypes significantly associated with milk or protein yield.....	54
Table 7: SNPs associated to calving ease from GWAS.....	71
Table 8: Additional SNPs sequenced in the LAP3 and NCAPG genes.....	73
Table 9: SNPs in the LAP3 gene.....	74

1 Foreword

This PhD thesis focuses on two key aspects of modern genomics, which can be imagined as the extremes of a complex technological and scientific pipeline: these are the *de novo* assembly of a genome and the use of dense Single Nucleotide Polymorphisms panels that results from the mapping of such markers on a reference genome sequence. Between these two extremes there is a wide number of other investigation areas such as genome annotation and SNP discovery that allows to exploit the existence of a reference genome into population-wide studies. This work results from the involvement in two research projects: the International Buffalo Genome Consortium and the national SELMOL project for the application of molecular methods to the improvement of domestic cattle and other species. The work presented summarizes my work in these wider framework. In particular, it covers the assembly of the water buffalo genome using 2nd generation sequencing techniques and the application of the Bovine BeadChip v.2 to the study of five major national cattle breeds.



International Buffalo Genome Consortium



SELMOL project

2 Theoretical background

2.1 Introduction

Each cell of a living organism contains chromosomes composed of a sequence of DNA base pairs. This sequence, the genome, represents a set of instructions that controls the replication and function of each organism[1]. The complete structure of an individual's or species genomic structure can be determined, within certain limits, using a combined set of laboratory and bioinformatics techniques called genome assembly. Genomics is the analytic and comparative study of genomes, has become possible with the advent of modern and automated techniques which allow scientists to decode entire genomes. Current technologies (2nd generation sequencing) allows to produce DNA molecules (reads) of length up to a few thousands of bases, orders of magnitude less than the total genomic length of a even a bacteria, with a cost proportional to actual length. To overcome this limitation, a technique called shotgun sequencing[2] has been developed (Figures 1-3):

1. the DNA sequence of a given organism is sheared into a large number of small fragments (*whole genome shotgun sequencing* or WGSS)
2. the ends of the fragments are sequenced
3. the resulting sequences are joined together using a computer program called an assembler

The basic assumption of this approach is that two reads sharing a common substring (i.e. the same sequence of nucleotides) will probably originate from the same genomic region; the assembler program exploits these overlaps to join the reads together as in a jigsaw puzzle[3]. However, there is a number of problems that must be taken into account when joining the reads: first the shearing problem is (not completely) random and thus to ensure reasonable covering of all DNA we have to sequence much more reads (from 8/10X to 50/60X more nucleotides than actual length) and even in this way there's still a portion of the genome that will be lost because of sequencing technology limitations; second, reads are short which means that overlaps are not unique and a number of branching overlaps is obtained when analysing the data; third genomes contain long repeated regions that may fool the sequencing algorithm. A number

of sequencing and algorithmic solutions has been proposed to address these problems that have lead to different assembly strategies. Shotgun sequencing was introduced by Frederick Sanger in 1982[4]. A quantum leap forward was made in 1995, when a team led by Craig Venter and Robert Fleischmann of The Institute for Genomic Research (TIGR) and Hamilton Smith of Johns Hopkins University used it on a large scale to sequence the 1.83 million base pair (Mbp) genome of the bacterium *Haemophilus influenzae*[5]

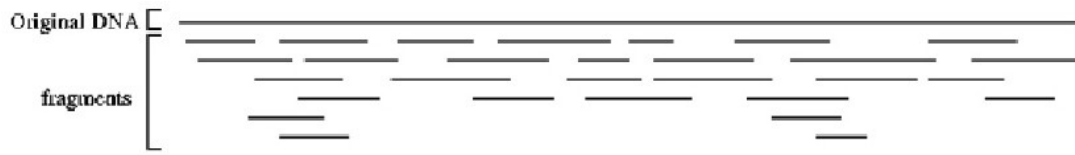


Figure 1: Genomic DNA is broken into a collection of fragment

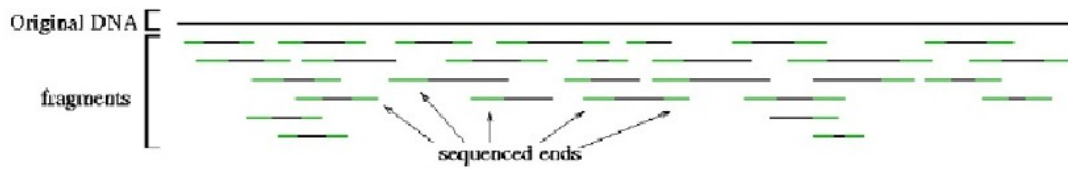


Figure 2: The ends of each fragment are sequenced

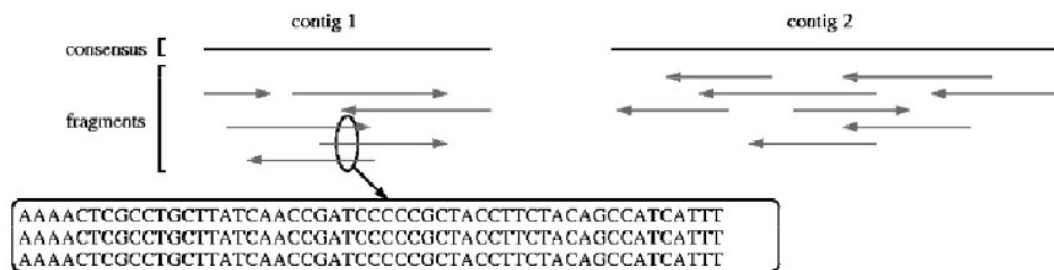


Figure 3: The sequence reads are assembled together based on sequence similarity

Even if the WGSS strategy was the base of these assembly projects it was applied at the end of a pipeline that involved the use of large DNA sequences called *Bacterial Artificial Chromosomes (BAC)*[6],[7]: DNA was broke into BACs which were then mapped to the genome to obtain a tiling path, after which the shotgun method is used to sequence each BAC in the tiling path separately. In contrast, whole-genome shotgun sequencing (WGSS) assembles the genome from the initial fragments without using a BAC map. The first genome assembled completely with the WGSS protocol was *Drosophila melanogaster* in 2000 by Myers and coworkers[8] and in 2001 the first version of the human genome was published by Venter and coworkers[9].

2.1.1 Sanger Sequencing

Sanger sequencing [4] was the first widespread method used for genome assembly and the first algorithmic strategies to assemble genomes were based on this technology. This method requires a first step (see Figure 4) in which is generated a certain amount of DNA to be sequenced. Starting from a single oligonucleotide of DNA named 'primer', it is possible to produce unlimited clones of the same fragment. This amplification is permitted by the use of DNA polymerase, an enzyme that allows the synthesis of a new strand of DNA because binds the deoxynucleotides using the primer as a template. One of deoxyribonucleotides triphosphate (dNTP) required for the synthesis of DNA is marked with radioactive markers. If radioactive nucleotides are present, they are randomly incorporated into the growing chain causing the chain termination. The cloned DNA can be divided into four aliquots in each using a different type of ddNTP molecules and giving rise to ending each with a different type of radioactive dideoxynucleotide. Within each tube will have millions of chains of variable length, but each with the same base at the 5' end, which is the primer. The products of four reactions must be loaded on a denaturing polyacrylamide gel, which is able to separate DNA strands that differ even a single nucleotide. At the end we will have four tracks each containing bands at different heights. The DNA chains all begin with the same primers, but are interrupted at different points. All the broken chains at the same point form a band, all those broken in the next step will form another distinct band from the previous, although the difference was only one single nucleotide. At the end of electrophoresis, the position of the radioactive DNA bands are determined by autoradiography from which is possible to view the sequence of the newly synthesized chain. Sanger method has been greatly improved since its first release in 1977, yet the cost time required to obtain a complete genome with this technology are still quite high.

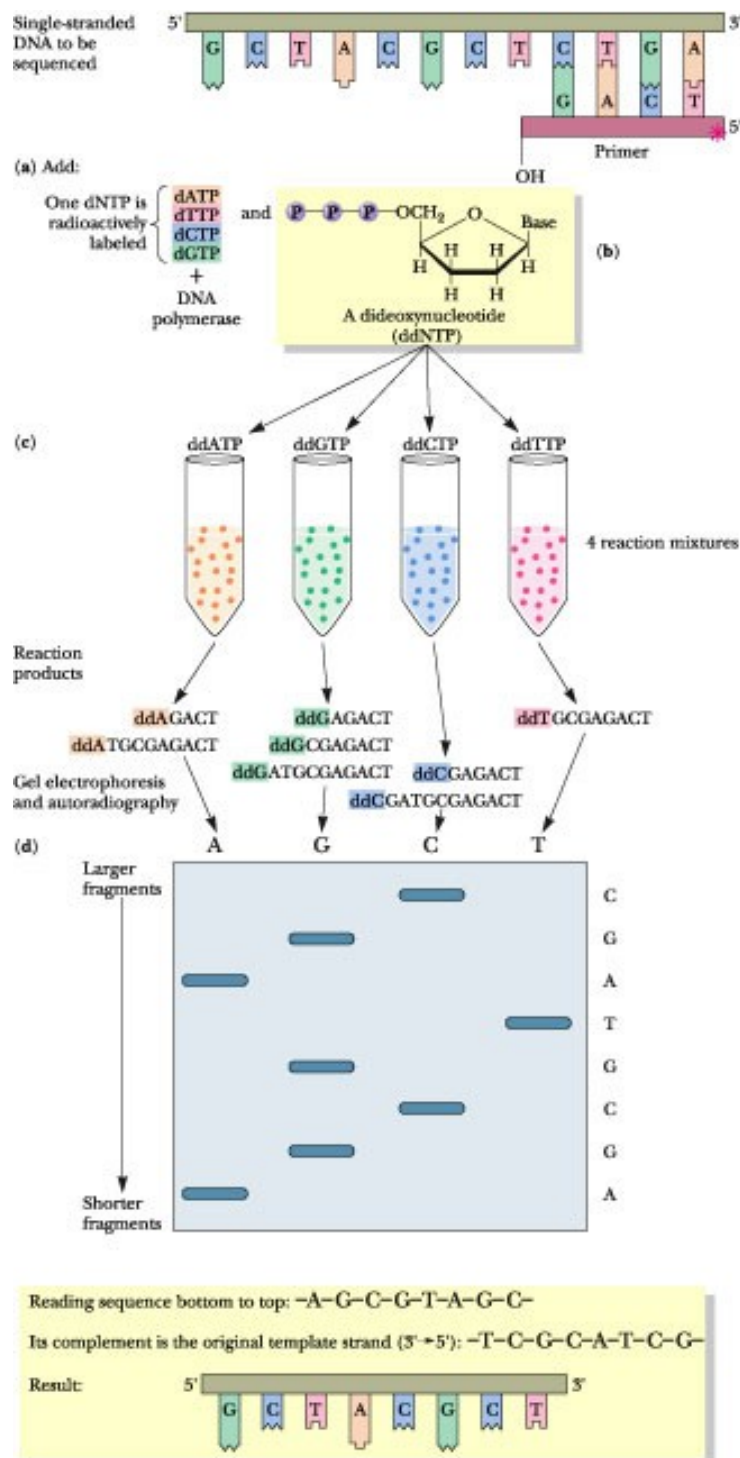


Figure 4: Sanger sequencing workflow

2.1.2 Second Generation Sequencing

The development of 2nd generation sequencing technologies [10], [11] allowed to overcome the cost limitations in sequencing experiments posed by the Sanger approach. With these techniques, developed starting from 2000 the cost per base sequenced was

reduced by orders of magnitude. Currently, three main technologies are used in genome assembly research: Roche 454, Illumina-SOLEXA and Applied Biosystems' SOLiD. All of these share common steps (see Figure 5) and reads produced by most 2nd sequencing technologies can be grouped into two wide categories:

1. single end reads, obtained sequencing one end of a DNA molecule
2. paired end and mate pair reads, obtained sequencing both ends of a DNA molecule, separated by a unsequenced region of known length (insert size) from 100/200 bp up to several Kbp; pairs are particularly useful to skip over repeat regions and to join different set of reads grouped by the assembler (contigs).
- 3.

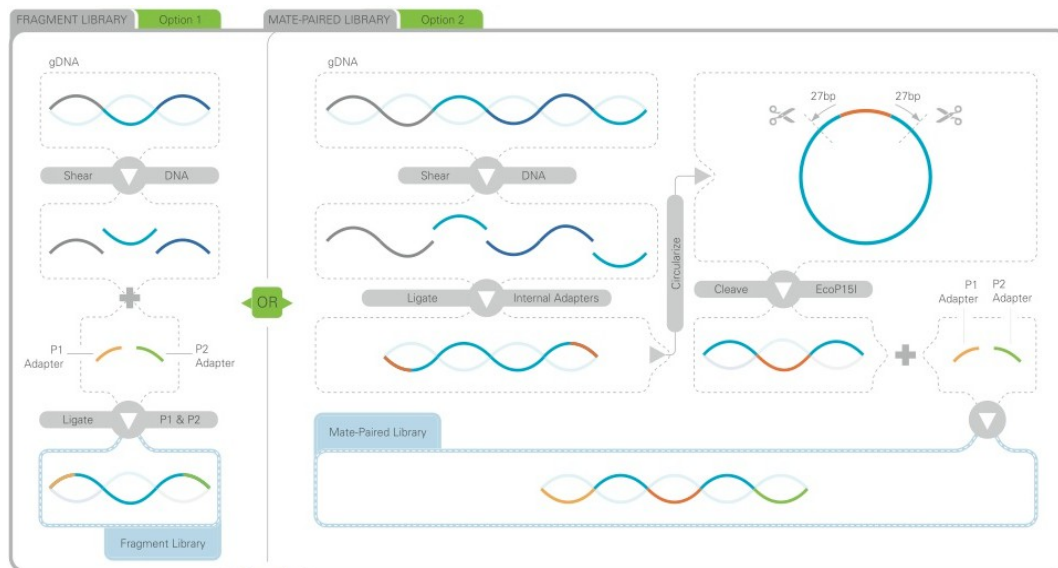


Figure 5: Second generation sequencing read types

2.1.2.1 Roche 454

454 sequencing [12] is a sequencing system, currently commercialized by Roche Applied Science corporation since 2004 (<http://www.454.com/>). The DNA to be sequenced is first fractionated into fragments (see Figure 6) which are fitted with adapter oligonucleotides on either end and denatured. These adapters provide priming sequences for both amplification and sequencing of the sample-library fragments. The DNA is fixed to the surface of streptavidin-coated beads on which is done a PCR reaction, with the addition of emulsion of water/oil to avoid cross contamination between the sequences belonging to different beads, and the result is a “DNA genotec” derived from each fragments. Each beads (DNA genotec) is then placed in a well of a special

chip optical fibers named 454 picotiter plate whose pores have a diameter of about 44 microns and is introduced a mix of polymerases, luciferase, apyrase and triphosphates to make the pyro-sequencing. The Pyrosequencing is a technology sequencing by synthesis. The technique allows to detect the bioluminescence produced at the end of a cascade of enzymatic reactions, which, in turn, is triggered by the incorporation of a nucleotide. The intensity of the light signal is directly proportional to the number of bases introduced by the polymerase in the new strand of DNA. The sulfurylase enzyme permits the conversion of the Ppi into ATP in the presence of adenosine 5' phosphosulfate. This ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin with the emission of photons in proportion to the amount of ATP. At each cycle the entire chip is scanned by a CCD which records light signals corresponding to bases inserted.

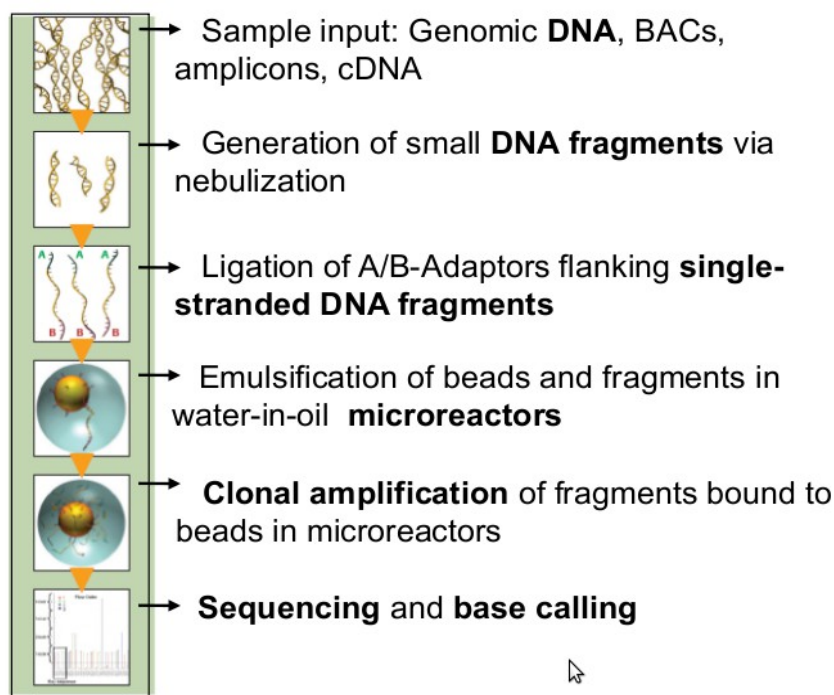


Figure 6: 454 sequencing workflow

2.1.2.2 Illumina

The SOLEXA technique, Figure 7, acquired by Illumina in 2006 for over 600 M\$ (www.illumina.com), consists of a proprietary flow cell surface for the parallel sequencing of DNA fragments (<http://seqanswers.com/forums/showthread.php?t=21>). Illumina reads are much shorter than 454 reads but their cost per base is significantly lower and their functional performance is higher. DNA is first fragmented, then it is at-

tached to adapters and denaturated in NaOH, then hybridization buffer is added to shift pH to neutral value and library is placed into the channel. In this way primer hybridize with a library molecule and the complementary strand have a chance to form a hybrid with an other primer. The fragments are then amplified locally and at each cycle the extended molecules bends and hybridizes to a second PCR primer forming a bridge. After more iterations the channel is characterized by clusters of oligonucleotides which are identical. At this point the system reads out the sequences of all clusters simultaneously.

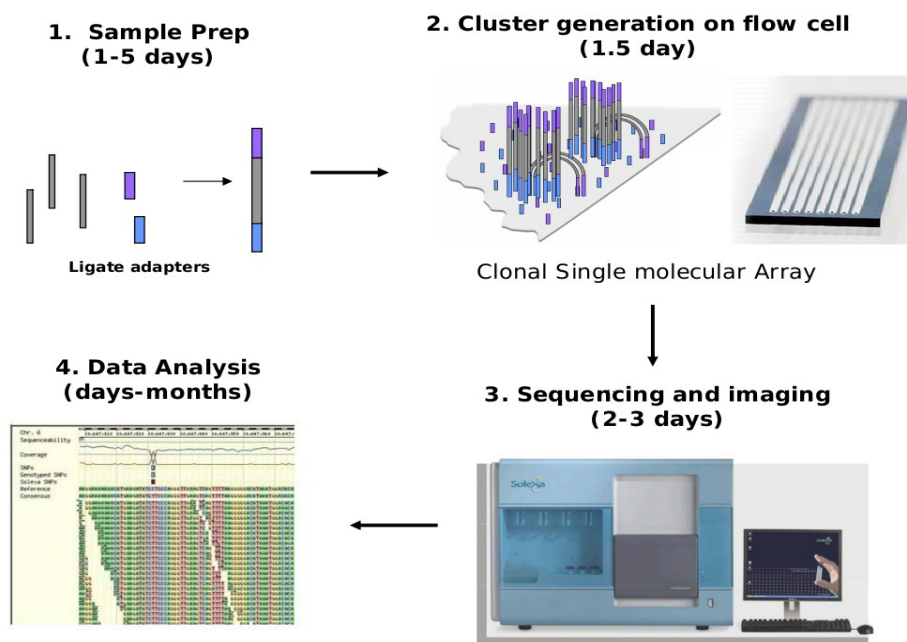


Figure 7: Illumina sequencing workflow

2.1.2.3 SOLiD

Sequencing by Oligonucleotide Ligation and Detection is a next-generation sequencing technology developed by Life Technologies (). This technique is based on “sequencing ligation” of fluorochoime-labeled oligonucleotides [13]. The DNA fragments hybridize to a specific section of the adapter. A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length. Following a series of ligation cycles, the extension

product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles.

2.1.2.4 Third Generation Sequencing

It took nearly two decades to go from the release of the first semi-automated genome sequencer in the mid-1980s to the launch of Roche's flagship 454 FLX next generation sequencer in 2005. The new wave of sequencers, (third generation sequencing) will likely enable scientists to reach the goal of the \$US1000 human genome. But to do this, the cost of the sequencing chemistry needs to drop to around \$US0.0000005/base, or two millionths of a dollar per base (at a 10x coverage). Currently, the cheapest 2nd generation sequencers cost about \$US0.000001/base. And while this is a gigantic improvement on 1985 prices when sequencing cost roughly \$US10/base, we are still well short of the price target.

Third generation [14–16] of sequencing technology sees single molecules of DNA being sequenced without the need for cloning or PCR amplification and the inherent biases these procedures introduce. There are generally two types of detection methods for single molecule sequencing: those that rely on fluorescence and CCD capture, and those that don't. Instruments that use the first of these detection methods include the Helicos Heliscope (<http://www.helicosbio.com/Products/HelicosregGeneticAnalysisSystem/HeliScopetradeSequencer/tabid/87/Default.aspx>), launched in 2008; Pacific Biosciences single molecule real time sequencing (SMRT) machines (<http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage>), which have been shipped to their first customers; and Life Technologies-VisiGen system, which relies on fluorescence resonance energy transfer (FRET).

2.1.3 De Novo assembly algorithms

2.1.3.1 Introduction

After DNA has been physically sheared into millions of reads, the *assembler* pieces together the many overlapping reads and reconstructs the original sequence. An assembled genome is usually composed of the following elements from simple to complex see Figure 8):

1. reads: the raw initial data
2. unitigs: unique merging of overlapping reads

3. contigs: merging of overlapping unitigs plus consensus sequence; uses read pairing
4. scaffolds: linear sequence of contigs joined by mate pairs; uses read pairing

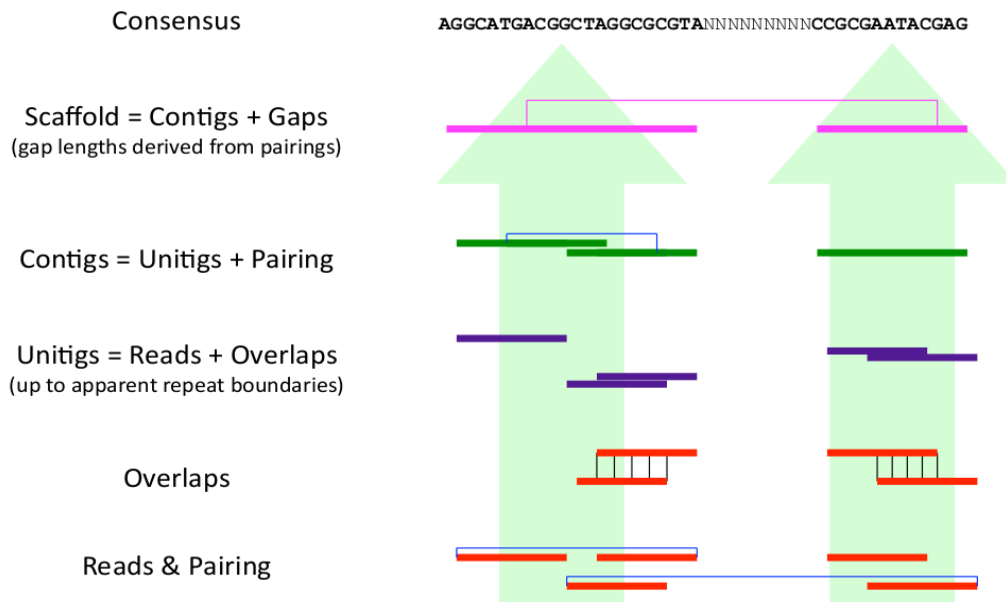


Figure 8: Elements of a DNA assembly

These elements are generated using data abstraction structures called graphs. The two most important graph type in current assembly packages are the overlap graph and the De Bruijn graph; use of either of them actually characterize the two main assembly algorithms i.e. Overlap-Layout-Consensus (OLC) and De Bruijn graph approach [17–19].

2.1.3.2 The Overlap-Layout-Consensus approach

This approach consist of computing all pair-wise overlaps between each reads and collects information into the overlap graph. Each node represents a read and an edge denotes an overlap between two reads. The overlap graph is used to compute a layout of reads and a consensus sequence of contigs. This method works best when there is a limited number of reads with significant overlap. A limit of this approach is that for the assembly of a simple organism the number of reads is very high (millions) and consequently the overlap graph became much large.

The OLC approach was typical of the Sanger-data assemblers. It was optimized for large genomes in software including Celera Assembler [8], Arachne [20], and CAP and PCAP [21]. OLC assemblers have three phases[22]:

1. Overlap discovery involves all-against-all, pair-wise read comparison.
2. Construction and manipulation of an overlap graph leads to an approximate read layout. The overlap graph need actual sequences, so large-genome graphs may fit into practical amounts of computer memory.
3. Multiple sequence alignment (MSA) determines the precise layout and then the consensus sequence. There is no known efficient method to compute the optimal MSA [23]. Therefore, the consensus phase uses progressive pair-wise alignments guided by, for instance, the approximate read layout. Sequences must be loaded in RAM but computations may be partitioned by contig.

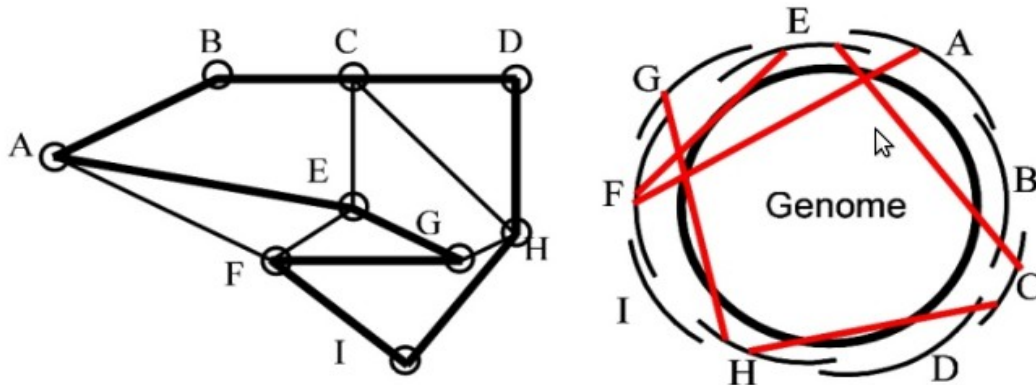


Figure 9: Overlap graph for a bacterial genome

The Celera Assembler is a Sanger-era OLC assembler revised for 454 data [24]. The revised pipeline, called CABOG, discovers overlaps using compressed seeds. CABOG reduces homopolymer runs, that is, repeats of single letters, to single bases to overcome homopolymer run length uncertainty in data. CABOG builds initial unitigs excluding reads that are substrings of other reads. Substrings account for a large portion of the data in high-coverage 454 data due to highly variable read lengths. CABOG avoids the substring-reads initially because they are more susceptible to repeat-induced false overlaps.

2.1.3.3 The De Bruijn graph approach

De Bruijn Graph (DBG) [25] can greatly reduce the computing resources necessary to assembly of a large number of reads by breaking reads into

smaller sequences of DNA, called k-mers (k parameter indicates the length of these small DNA sequences). This method permits to collect overlap information between the k-mers created rather than the reads. When two k-mers overlap by k-1 length, there is a directed edge from the first k-mer to the succeeding k-mer with a corresponding edge in the reverse direction. Multiple edges are derived if one k-mer overlaps with multiple, different k-mers by k-1 length on the side. A k-mer is considered to differ only in orientation with its reverse complement, and only one of the two is saved. Assembly packages based on the DBG include Velvet, Euler, Abyss and SOAP-denovo [26–28].

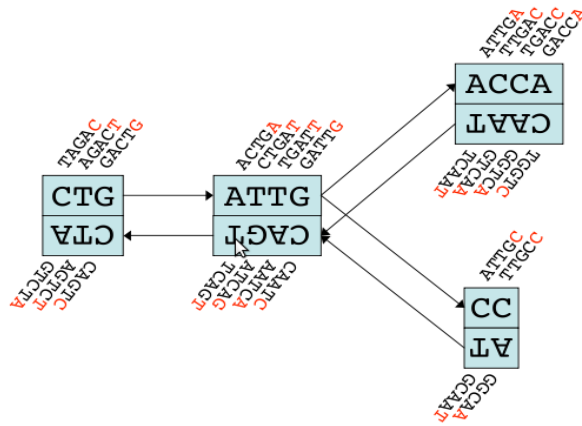


Figure 10: Schematic diagram of the de Bruijn graph implementation

Four factors complicate the application of K-mer graphs to DNA sequence assembly:

1. DNA is double stranded. The forward sequence of any given read may overlap the forward or reverse complement sequence of other reads and several different implementations of the DBG have proposed to address this problem.
2. Real genomes present complex repeat structures including tandem repeats, inverted repeats, imperfect repeats, and repeats inserted within repeats. Repeats longer than K lead to tangled K-mer graphs that complicate the assembly problem.
3. A palindrome is a DNA sequence that is its own reverse complement. Palindromes induce paths that fold back on themselves.
4. Real data includes sequencing error. DBG assemblers use several techniques to reduce sensitivity to this problem. First, they pre-process the reads to remove error. Second, they weight the graph edges by the number of reads that support

them, and then “erode” the lightly supported paths. Third, they convert paths to sequences and use sequence alignment algorithms to collapse nearly identical paths.

2.1.3.4 The MSR-CA assembler

To overcome the limits of both OLC and the De Bruijn approaches Zimin et al. (Zimin et al., unpublished data) proposed a method that uses the De Bruijn graph to merge 2nd generation reads into longer reads (Super-Reads) and thus obtain a lossless data set dimension shrinkage making it possible to apply the OLC method. Super-Reads are created mapping reads on the De Bruijn graph and then merging them when a unique path is detected between k-mers thus yielding (almost) zero information loss. The MSR-CA assembler pipeline is composed of three main steps (see Figure 12) which corresponds to the three main software components:

1. K-mer counting performed with the Jellyfish package
2. Error-correction and Super-Reads creation using the De Bruijn Graph; this is performed with software developed by Zimin et al.
3. Genome assembly using Super-Reads performed with a modified version of the OLC CABOG Assembler

Jellyfish generates k-mer counting statistics and is highly memory intensive; a mammalian data with $1-2 \times 10^9$ reads may require 256GB of RAM memory. Error-correction attempts to remove sequencing errors using k-mer frequency; when correction is not possible reads are trimmed or completely discarded. Corrected reads generate much less improper branching in the graph and lead to better assembly results. Super-Reads have been covered above and CABOG is the modified version of the Celera assembler able to work with 2nd generation data.

The basic goal of MSR-CA is to reduce the complexity of the overlap-based assembly by transforming the high coverage (>50x) of the short reads into 5-6x coverage by super-reads, while making only the “easy” assembly decisions. This is done by creating a DBG is built using the most abundant data available (usually Illumina short paired ends) and then counting how many times each k-mer occurs in our reads. Each k-mer may be extended either by a A,C,G, or T (see Figure 11). If only one of these four occurs in the table, we say it has a unique following (or preceding) k-mer (such a k-mer is a *simple* k-mer). Examining the two simple (if any) k-mers at the ends of a read the read is extended by one base, using the uniquely occurring extension. Now the exten-

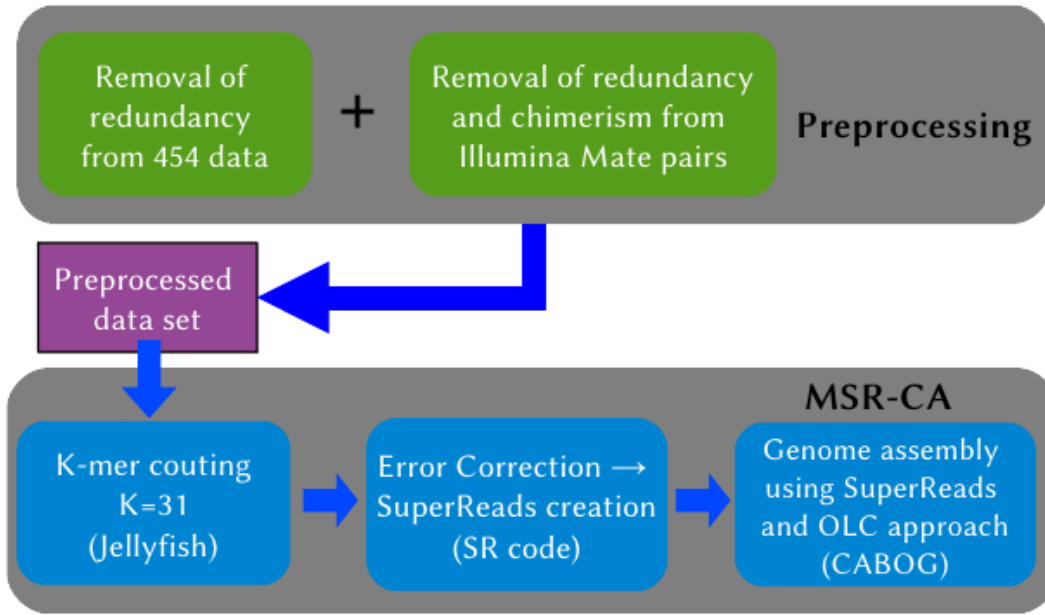


Figure 11: MSR-CA workflow

ded read has a new k-mer on its end. The process is repeated on each end until the k-mers on the end are no longer simple. The process stops at either end when the extension is not unique or if there is indeed no extension. The resulting read is defined as a *Super-read*. The two most important properties of the *Super-reads* are:

1. by design they are at least as long as the original reads
2. many of the original reads yield the same super-read. Using super-reads leads to vastly reduced data set

This much smaller data set is then assembled with a OLC assembler.

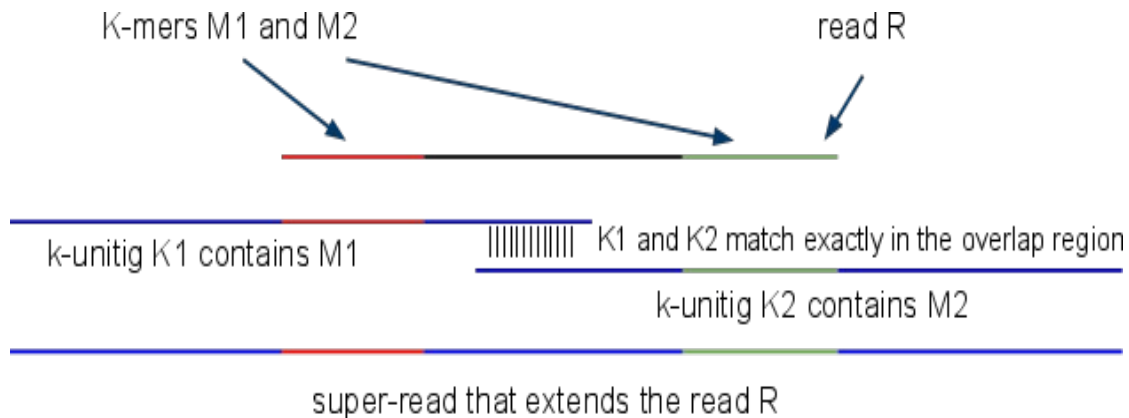


Figure 12: An example of a read whose super read has two k-unitigs

Read R contains k-mers M1 and M2 on its ends. M1 and M2 each belong to one and only one k-unitig K1 and K2. K-unitigs K1 and K2 are shown in blue and the matching k-mers M1 and M2 are shown in red and green respectively. K1 and K2 overlap by $k-1$ bases. We extend read R on both ends producing the super-read also depicted in blue. A super-read can consist of one k-unitig or can contain many k-unitigs.

2.2 SNP panels

2.2.1 Introduction

Single Nucleotide Polymorphisms (SNPs) are the most common form of genetic variation between individuals, SNPs occur approximately once every 1,000 bases in mammals. SNPs are DNA sequence variations that occur when a single nucleotide (A,T,C, or G) in the genome sequence is altered (see Figure 13). Millions of these variants are indexed in the National Center for Biotechnology Information's dbSNP database (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), covering several organisms. To be classified as a SNP a variation's Minimum Allele Frequency (MAF) must be more than 1% (i.e. the rare variant needs to be present in at least 1% of the studied population) and must be inherited. On the other hand if the variation arises in a given generation or is less common is called a mutation.

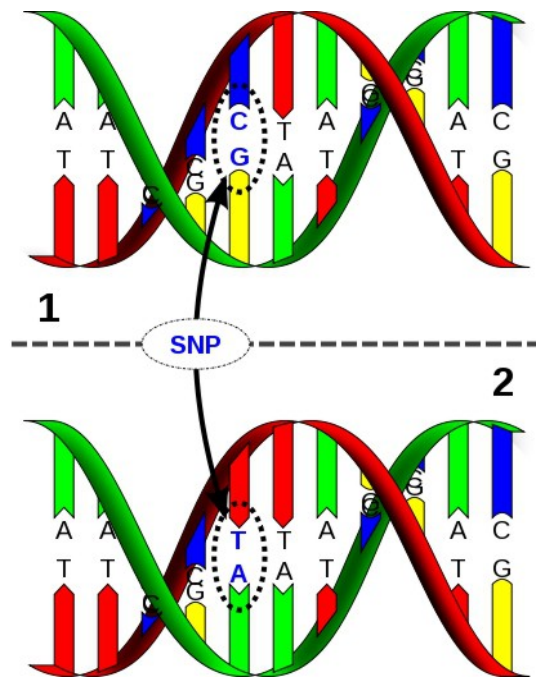


Figure 13: Single Nucleotide Polymorphisms

SNP arrays (http://en.wikipedia.org/wiki/SNP_array) are a type of DNA microarrays used to detect polymorphisms within a population comparing the studied sample with a reference genome and are thus species-specific. Microarray technology is well suited to SNP detection. Each SNP allele can be encoded in a single 25-base (or 50-

base) oligonucleotide, and genotypes can be assigned with extremely high accuracy (99.8%; [29]) by detection of hybridization differences in genomic DNA (heterozygotes and homozygotes can also be identified). Large numbers of SNP-specific oligonucleotides can be placed on a single chip and assayed in parallel on a single DNA sample, so the overall cost per genotype becomes very low. This technology allows a set of samples to be profiled for SNPs that cover the entire genome at high resolution, a "whole-genome association study." There are approximately 10 million SNPs across the worldwide human population, and over 5 million of these have sequences available in dbSNP at the NCBI Website (www.ncbi.nlm.nih.gov/SNP/snpsummary.cgi). Affymetrix currently produces arrays of 10,000, 100,000, 500,000, and 1 million human SNPs. Using a different technology, Illumina offers a HumanHap300-Duo Genotyping BeadChip that contains a total of about 1 million tag SNPs from the HapMap Project. SNPs used to study other species such as cattle or sheep are usually less dense and contain $\sim 10^5$ although more dense panels are being developed for species of economic importance.

2.2.2 The Illumina BovineSNP50 chip

The BovineSNP50 v1 BeadChip (Figure 14) used in this work contains 54,001 SNPs uniformly distributed across the cattle genome (http://www.illumina.com/products/bovine_snp50_whole-genome_genotyping_kits.ilmn). It was developed by Illumina in collaboration with the USDA-ARS, University of Missouri, and the University of Alberta. More than 24,000 SNP probes target novel SNP loci that were discovered by sequencing three pooled populations of economically important beef and dairy cattle using Illumina's Genome Analyzer. Additional content was derived from publicly available sources such as the bovine reference genome, Btau1, and the Bovine HapMap Consortium data set. The current version (v2) contains 54,609 SNPs. Recently, Illumina has presented the 777,000 SNP BovineHD BeadChip which expands the diversity of bovine breeds assessed in genetic prediction and enables more discoveries of quantitative traits.



Figure 14: Illumina SNP50 Beadchip

2.3 Statistical methods

2.3.1 Infinitesimal model and Quantitative Trait Locus model (QTL)

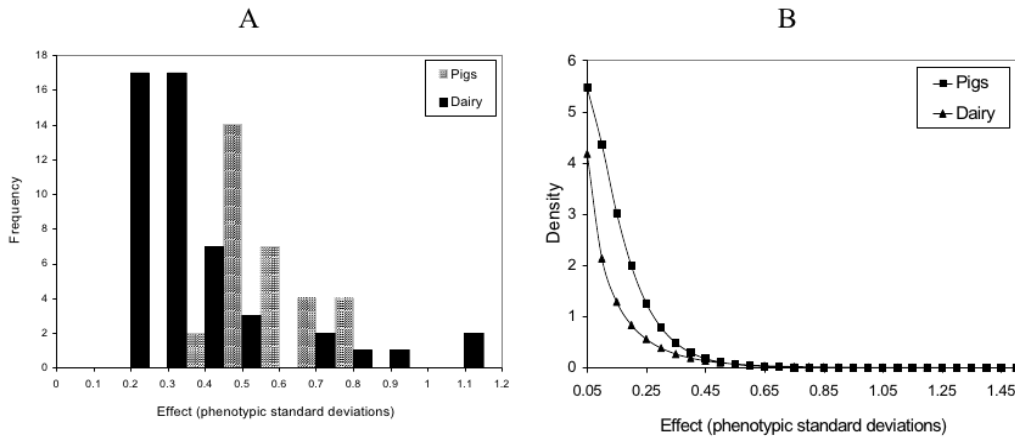


Figure 15: Distribution of additive (QTL) effects from pig experiments

Distributions are scaled by the standard deviation of the relevant trait, and distribution of gene substitution (QTL) effects from dairy experiments scaled by the standard deviation of the relevant trait. B. Gamma Distribution of QTL effect from pig and dairy experiments, fitted with maximum likelihood.

The vast majority of economically important traits in livestock and aquaculture production systems are quantitative, that is they show continuous distributions. In attempting to explain the genetic variation observed in such traits, two models have been proposed, the infinitesimal model and the finite loci model. The infinitesimal model assumes that traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect [30]. This model has been exceptionally valuable for animal breeding, and forms the basis for breeding value estimation theory (e. g. [31]). However, the existence of a finite amount of genetically inherited material (the genome) and the revelation that there are perhaps a total of only around 20,000 genes or loci in the genome [32], means that there is must be some finite number of loci underlying the variation in quantitative traits. In fact there is increasing evidence that the distribution of the effect of these loci on quantitative traits is such that there are a few

genes with large effect, and a many of small effect ([33], [34]). In Figure 15, the size of quantitative trait loci (QTL) reported in QTL mapping experiments in both pigs and dairy cattle is shown. These histograms are not the true distribution of QTL effects however, they are only able to observe effects above a certain size determined by the amount of environmental noise, and the effects are estimated with error. In Figure 15A, the distribution of effects adjusted for both these factors is displayed. The distributions in Figure 15B indicate there are many genes of small effect, and few of large effect. The search for these loci, particularly those of moderate to large effect, and the use of this information to increase the accuracy of selecting genetically superior animals, has been the motivation for intensive research efforts in the last two decades. Two approaches have been used to uncover QTL. The candidate gene approach assumes that a gene involved in the physiology of the trait could harbour a mutation causing variation in that trait. The gene, or parts of the gene, are sequenced in a number of different animals, and any variations in the DNA sequences, that are found, are tested for association with variation in the phenotypic trait. This approach has had some successes – for example a mutation was discovered in the oestrogen receptor locus (ESR) which results in increased litter size in pigs [35],[36] There are two problems with the candidate gene approach, however. Firstly, there are usually a large number of candidate genes affecting a trait, so many genes must be sequenced in several animals and many association studies so many genes must be sequenced in several animals and many association studies carried out in a large sample of animals (the likelihood that the mutation may occur in non-coding DNA further increases the amount of sequencing required and the cost). Secondly, the causative mutation may lie in a gene that would not have been regarded a priori as an obvious candidate for this particular trait. An alternative is the QTL mapping approach, in which chromosome regions associated with variation in phenotypic traits are identified. QTL mapping assumes the actual genes which affect a quantitative trait are not known. Instead, this approach uses neutral DNA markers and looks for associations between allele variation at the marker and variation in quantitative traits. A DNA marker is an identifiable physical location on a chromosome whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or more often some segment of DNA with no known coding function but whose pattern of inheritance can be determined. When DNA markers are available, they can be used to determine if variation at the molecular level (allelic variation at

marker loci along the linkage map) is linked to variation in the quantitative trait. If this is the case, then the marker is linked to, or on the same chromosome as the investigated locus, a quantitative trait locus or QTL which has allelic variants causing variation in the quantitative trait. Until recently, the number of DNA markers identified in livestock genome was comparatively limited, and the cost of genotyping the markers was high. This constrained experiments designed to detect QTL to using a linkage mapping approach. If a limited number of markers per chromosome are available, then the association between the markers and the QTL will persist only within families and only for a limited number of generations, due to recombination. For example in one sire, the A allele at a particular marker may be associated with the increasing allele of the QTL, while in another sire, the a allele at the same marker may be associated with the increasing allele at the QTL, due to historical recombination between the marker and the QTL in the ancestors of the two sires. To illustrate the principle of QTL mapping exploiting linkage, consider an example where a particular sire has a large number of progeny. The parent and the progeny are genotyped for a particular marker. At this marker, the sire carries the marker alleles 172 and 184, Figure 16. The progeny can then be sorted into two groups, those that receive allele 172 and those that receive allele 184 from the parent. If there is a significant difference between the two groups of progeny, then this is evidence that there is a QTL linked to that marker.

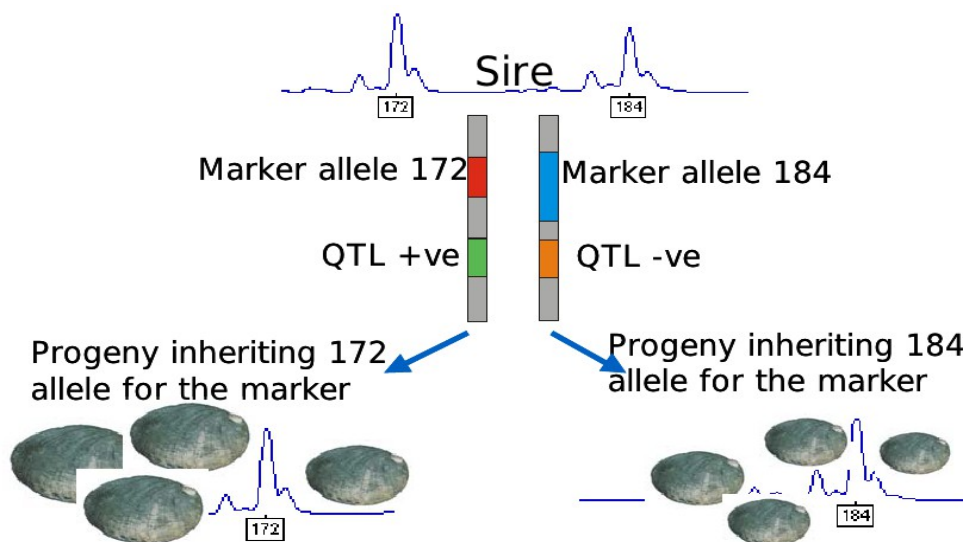


Figure 16: Linkage mapping

Principle of quantitative trait loci (QTL) detection, illustrated using an abalone example. A sire is heterozygous for a marker locus, and carries the alleles 172 and 184 at this locus. The sire has a large number of progeny. The progeny are separated into two groups, those that receive allele 172 and those that receive allele 184. The significant difference in the trait of average size between the two groups of progeny indicates a QTL linked to the marker. In this case, the QTL allele increasing size is linked to the 172 allele and the QTL allele decreasing size is linked to the 184 allele.

QTL mapping exploiting linkage has been performed in all nearly livestock species for a huge range of traits [36]. The problem with mapping QTL exploiting linkage is that, unless a huge number of progeny per family or half sib family are used, the QTL are mapped to very large confidence intervals on the chromosome. To illustrate this, consider this formula [37] for estimating the 95% CI for QTL location for simple QTL mapping designs under the assumption of a high density genetic map:

$$CI=3000/(kN\delta^2)$$

$$CI=\frac{3000}{k N \delta^2} \quad (1)$$

where N is the number of individuals genotyped, δ allele substitution effect (the effect of getting an extra copy of the increasing QTL allele) in units of the residual standard deviation, k the number of informative parents per individual, which is equal 1 for half-sibs and backcross designs and 2 for F2 progeny, and 3000 is about the size of the cattle genome in centi-Morgans. For example, given a QTL segregates on a particular chromosome within a half sib family of 1000 individuals, for a QTL with an allele substitution effect of 0.5 residual standard deviations the 95% CI would be 12 cM. Such large confidence intervals have two problems. Firstly if the aim of the QTL mapping experiment is to identify the mutation underlying the QTL effect, in a such a large interval there are a large number of genes to be investigated (80 on average with 20,000 genes and a genome of 3000 cM). Secondly, use of the QTL in marker assisted selection is complicated by the fact that the linkage between the markers and QTL is not sufficiently close to ensure that marker-QTL allele relationships persist across the population, rather marker-QTL phase within each family must be established to implement marker assisted selection. An alternative, if dense markers were available, would be to exploit linkage disequilibrium (LD) to map QTL. Performing experiments to map QTL in genome wide scans using LD has recently become possible due to the availability of tenths of thousands of single nucleotide polymorphism (SNP markers) in cattle, pigs, chickens and sheep in the near future. A SNP marker is a difference in nucleotide between animals (or an animals pair of chromosomes), at a defined position in the genome, e. g.:

Animal 1. ACTCGGGC

Animal 2. ACTTGGGC

Rapid developments in SNP genotyping technology now allow genotyping of a SNP marker in an individual for as little as 1c US.

2.3.2 Linkage Disequilibrium

The classical definition of linkage disequilibrium (LD) refers to the non-random association of alleles between two loci. Consider two markers, A and B, that are on the same chromosome. A has alleles A1 and A2, and B has alleles B1 and B2. Four haplotypes of markers are possible A1_B1, A1_B2, A2_B1 and A2_B2. If the frequencies of alleles A1, A2, B1 and B2 in the population are all 0.5, then we would expect the frequencies of each of the four haplotypes in the population to be 0.25. Any deviation of the haplotype frequencies from 0.25 is linkage disequilibrium (LD), i.e. the genes are not in random association. As an aside, this definition serves to illustrate that the distinction between linkage and linkage disequilibrium mapping is somewhat artificial – in fact linkage disequilibrium between a marker and a QTL is required if the QTL is to be detected in either sort of analysis. The difference is:

linkage analysis only considers the linkage disequilibrium that exists within families, which can extend for 10s of cM, and is broken down by recombination after only a few generations

linkage disequilibrium (see Figure 17) mapping requires a marker to be in LD with a QTL across the entire population. To be a property of the whole population, the association must have persisted for a considerable number of generations, so the marker(s) and QTL must therefore be closely linked.

One measure of LD is D, calculated as [38]:

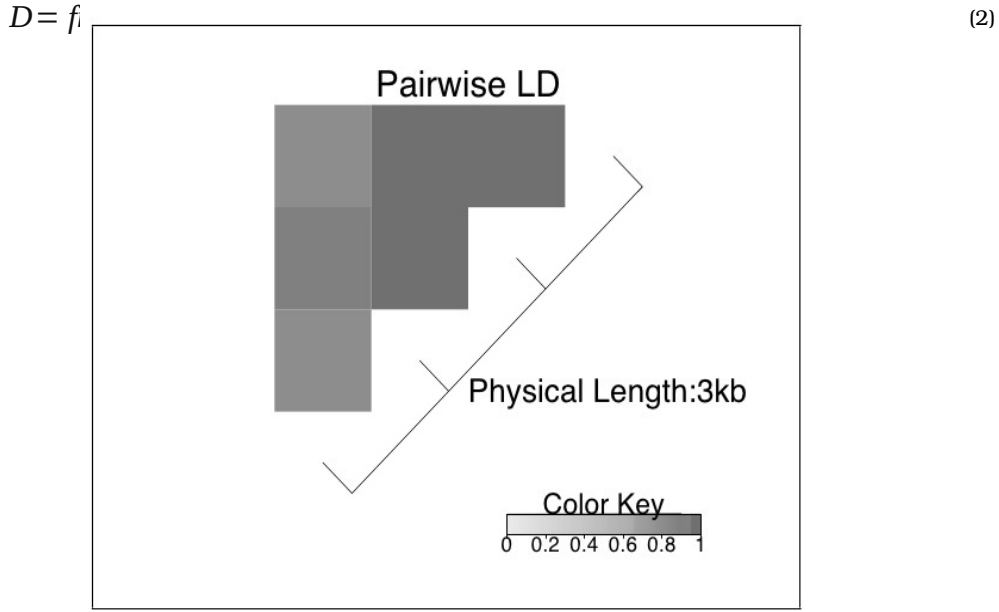


Figure 17: Pairwise Linkage Disequilibrium

where $\text{freq}(A1_B1)$ is the frequency of the A1_B1 haplotype in the population, and likewise for the other haplotypes. The D statistic is very dependent on the frequencies of the individual alleles, and so is not particularly useful for comparing the extent of LD among multiple pairs of loci (e. g. at different points along the genome). Hill and Robertson [39] proposed a statistic, r^2 , which was less dependent on allele frequencies:

$$r^2 = D / [\text{freq}(A1) \text{freq}(B1) \text{freq}(A2) \text{freq}(B2)] \quad (3)$$

where $\text{freq}(A1)$ is the frequency of the A1 allele in the population, and likewise for the other alleles in the population. Values of r^2 range from 0, for a pair of loci with no linkage disequilibrium between them, to 1 for a pair of loci in complete LD.

More precise estimates of LD are also tenable through fine mapping studies. Ultimately, through characterizing regions of high LD, mammalian genomes can be divided into LD blocks. These blocks are separated by hotspots, regions in which recombination events are more likely to occur [40]. An illustration of two LD blocks separated by a recombination hotspot is given in Figure 3.2. In general, alleles tend to be more correlated within LD blocks than across LD blocks. Once regions of high LD are identified, investigators aim to determine the smallest subset of SNPs that characterizes the variability in this region, a process referred to as SNP tagging. Here the goal is to reduce redundancies in the genetic data. Consider a pair of SNPs that are in perfect LD so that r^2 is equal to 1. Genotyping both SNPs in this case is unnecessary since the relationship between the two is deterministic. That is, by the definition of LD, know-

ledge of the genotype of one SNP completely defines the genotype of the second, and so there is no reason to sequence both loci. Well-defined tag SNPs will capture a substantial majority of the genetic variability within an LD block.

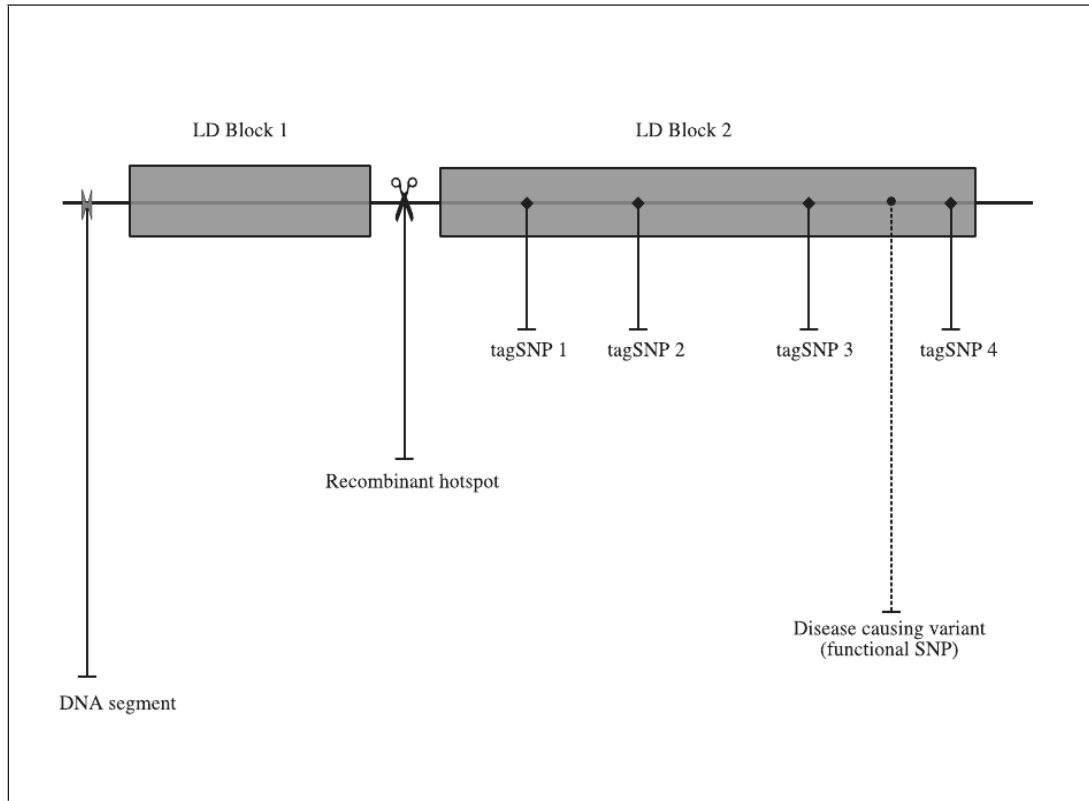


Figure 18: Illustration of LD blocks and associated tag SNPs

When investigating cattle breeds, QTL effects have generally been related to polymorphisms within functional candidate or other genes located in the QTL confidence interval. In an association study, a given SNP genotype may result associated to a phenotype if i) the SNP is the causative mutation; ii) the SNP is in high linkage disequilibrium with the gene under investigation; iii) the signal is a false positive (e.g. given by chance, by sampling bias due to a genetic sub-structure in the dataset, etc). For this reason only in a few cases causal mutations influencing milk traits have been identified using this approach; examples include K-casein ([41],[42]), SCD ([43],[44], DGAT1 ([45] and few others [46]. Even when there is a clear evidence of causality (e.g., DGAT1), genetic heterogeneity and genetic background may have a great impact on allele substitution effects [47]. Therefore, the potential application of gene/trait associations in breeding must be tested on a breed by breed base and evaluated with attention and prudence.

2.3.3 Population genetics

2.3.3.1 Introduction

Population genetics [48] is the study of the frequency of occurrence of alleles within and between populations. Frequency information can be applied to a variety of population issues, including understanding the genetic basis of traits of interest, developing breeding strategies, and investigating the evolutionary history of populations. Traditionally, the study of population genetics involved the identification of different alleles through observation of the expressed traits, broadly called the *phenotype*. Mendelian genetics allowed population geneticists to identify the heritable form of a gene (*genotype*) including individual variants (alleles). Advances in molecular genetics facilitated identification of single genes at the molecular level. Regardless of the method used to identify genes and alleles, population geneticists use statistics of allele frequencies to understand and make predictions about gene flow in populations - past, present, and future.

It is often said that variation in genes is necessary to allow organisms to adapt to ever-changing environments. However, it is actually the variation in alleles that is critical. Alleles are different versions of the same gene that are expressed as different phenotypes. New alleles appear in a population by the random and natural process of *mutation*, and the frequency of occurrence of an allele changes regularly as a result of *mutation*, *genetic drift*, and *selection*. Nearly all mammals are diploid [49], i.e. they have two homologous copies of each chromosome and thus two copies (two alleles) of each gene, one inherited from each parent. If an individual has two different versions of a particular gene, the individual is said to be heterozygous for that gene; if the two alleles are the same, the individual is homozygous. Since a population needs variation, the measure of the amount of heterozygosity across all genes can be used as a general indicator of the amount of genetic variability and genetic health of a population.

Observed Heterozygosity (H_o) is the proportion of genes that are heterozygous and the number of individuals that are heterozygous for each particular gene. For a single gene locus with two alleles H_o is:

$$H_o = \frac{n_{\text{heterozygotes}}}{\text{sample size}} \quad (4)$$

Derivations of the above formula are used to calculate the H_o when there are more than two alleles for a particular locus, which is particularly common when microsatellite or simple sequence repeat (SSR) markers are applied for analysis of populations.

Expected Heterozygosity: The Expected Heterozygosity (H_e) is defined as the estimated fraction of all individuals that would be heterozygous for any randomly chosen locus. The H_e differs from the H_o because it is a prediction based on the known allele frequency from a sample of individuals.

Deviation of the observed from the expected can be used as an indicator of important population dynamics. Based on Mendelian genetics, it is possible to predict the probability of the appearance of a particular allele in an offspring when the alleles of each parent are known. Similar predictions can be made about the frequencies of alleles in the next generation of an entire population. By comparing the predicted or "expected" frequencies with the actual or "observed" frequencies in a real population, one can infer a number of possible external factors that may be influencing the genetic structure of the population (such as inbreeding or selection).

2.3.3.2 Population structure

A population may be considered as a single unit. However, in many species and circumstances, populations are subdivided in smaller units. Such subdivision may be the result of ecological (habitats are not continuous) or behavioural factors (conscious or unconscious relocation). If a population is subdivided, the genetic links among its parts may differ, depending on the real degree of gene flow taking place.

A population is considered structured if (1) genetic drift is occurring in some of its subpopulations, (2) migration does not happen uniformly throughout the population, or (3) mating is not random throughout the population. A population's structure affects the extent of genetic variation and its patterns of distribution.

Here follows a little terminology in population genetics: ‘

Genetic drift’ refers to fluctuations in allele frequencies that occur by chance (particularly in small populations) as a result of random sampling among gametes. Drift decreases diversity within a population because it tends to cause the loss of rare alleles, reducing the overall number of alleles.

‘*Gene flow*’ is the passage and establishment of genes typical of one population in the genepool of another by natural or artificial hybridization and backcrossing.

'*Non-random mating*' occurs when individuals that are more closely (inbreeding) or less closely related mate more often than would be expected by chance for the population. Self-pollination or inbreeding is similar to mating between relatives. It increases the homozygosity of a population and its effect is generalized for all alleles. Inbreeding per se does not change the allelic frequencies but, over time, it leads to homozygosity by slowly increasing the two homozygous classes.

'*Mutations*' could lead to occurrence of new alleles, which may be favourable or deleterious to the individual's ability to survive. If changes are advantageous, then the new alleles will tend to prevail by being selected in the population. The effect of *selection* on diversity may be: (i) 'Directional', where it decreases diversity; (ii) 'Balancing', where it increases diversity. Heterozygotes have the highest fitness, so selection favours the maintenance of multiple alleles; and (iii) 'Frequency dependent', where it increases diversity. Fitness is a function of allele or genotype frequency and changes over time.

'*Migration*' implies not only the movement of individuals into new populations but that this movement introduces new alleles into the population (gene flow). Changes in gene frequencies will occur through migration either because more copies of an allele already present will be brought in or because a new allele arrives. The immediate effect of migration is to increase a population's genetic variability and, as such, helps increase the possibilities of that population to withstand environmental changes. Migration also helps blend populations and prevent their divergence.

Hardy-Weinberg Principle: Based on Mendel's principles of inheritance, G.H. Hardy and Wilhelm Weinberg, independently developed the concept that is known today as the 'Hardy-Weinberg Equilibrium' or 'Hardy-Weinberg Principle' ([50], [51]), which states: "In a large, randomly breeding (diploid) population, allelic frequencies will remain the same from generation to generation; assuming no unbalanced mutation, gene migration, selection or genetic drift." When a population meets all of the Hardy-Weinberg conditions it is said to be in Hardy-Weinberg equilibrium. This equilibrium can be mathematically expressed based on simple binomial (for two alleles) or multinomial (multiple allele) distribution of the gene frequencies.

Testing for Hardy-Weinberg Equilibrium: Populations in their natural environment can never meet all of the conditions required to achieve Hardy-Weinberg equilibrium, thus their allele frequencies will change from one generation to the next and the population

will evolve. Just how far the population deviates from Hardy-Weinberg is an indication of the intensity of external factors, and can be determined by a statistical formula called a chi-square, which is used to compare observed versus expected outcomes.

Effective Population Size: One of the many variables of population dynamics that can influence the rate and size of fluctuation in allele frequencies is population size. Genetic drift, the random increase or decrease of an allele's frequency, affects small populations more severely than large ones, since alleles are drawn from a smaller parental gene pool. The rate of change in allele frequencies in a population is determined by the population's effective population size. The effective population size is the number of individuals that evenly contribute to the gene pool.

The actual number of individuals in a population is rarely the effective population size. This is because some individuals reproduce at a higher rate than others (have a higher fitness), the distribution of males and females may result in some individuals being unable to secure a mate, or inbreeding reduces the unique contribution of an individual. The effective population size is a theoretical measure that compares a population's genetic behavior to the behavior of an "ideal" population. As the effective population size becomes smaller, the chance that allele frequencies will shift due to chance (drift) alone becomes greater.

Inbreeding and Relatedness: Small effective population size can result in a high occurrence of inbreeding, or mating between close relatives. One of the effects of inbreeding is a decrease in the heterozygosity (increase in homozygosity) of the population as a whole, which means a decrease in the number of heterozygous genes in the individuals. This effect places individuals and the population at a greater risk from homozygous recessive diseases that result from inheriting a copy of the same recessive allele from both parents. The impact of accumulating deleterious homozygous traits is called 'inbreeding depression' - the loss in population vigour due to loss in genetic variability or genetic options.

2.3.3.3 F statistics

In the 1950s, Sewall Wright [52] developed a set of parameters called *F statistics*. If we assume that genotypes in the base population⁴ were in Hardy-Weinberg proportions, then F_{IS} is the probability that two alleles in an individual are identical by descent (relative to the subpopulation from which they are drawn), F_{ST} is the probability that two alleles drawn at random are identical by descent (relative to the combined

populations), and F_{IT} is the probability that two alleles in an individual are identical by descent (relative to the combined population). A random variable indicating whether or not two alleles are identical by descent takes on only two values, so these probabilities are equivalent to a correlation. The relationships among the F statistics can be deduced through the following:

$$F_{IT} = 1 - \frac{H_I}{H_T} \quad (5)$$

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$$

where, H_T = total gene diversity or expected heterozygosity in the total population as estimated from the pooled allele frequencies, H_I = intrapopulation gene diversity or average observed heterozygosity in a group of populations, and H_S = average expected heterozygosity estimated from each subpopulation. F statistics estimate the deficiency or excess of average heterozygotes in each population (F_{IS}), the degree of gene differentiation among populations in terms of allele frequencies (F_{ST}) and the deficiency or excess of average heterozygotes in a group of populations (F_{IT}). There are two sources of allele frequency difference among subpopulations in our sample:

- (1) real differences in the allele frequencies among our sampled subpopulations and
- (2) differences that arise because allele frequencies in our samples differ from those in the subpopulations from which they were taken.

Nei and Chesser [53] described the G_{ST} approach to account for the sampling error. G_{ST} is an interpopulation differentiation measure when multiple loci are used for analysis. In other words, it measures the proportion of gene diversity that is measured among populations, when a large number of loci are sampled.

$$G_{ST} = \frac{D_{ST}}{H_T} \quad (6)$$

$$H_T = H_S + D_{ST}$$

$$D_{ST} = H_T - H_S$$

where, D_{ST} is the interpopulation diversity, H_T the total diversity and H_S the intrapopulation genic diversity.

2.3.4 Genome-wide scan (GWAS) using SNP panels

The introduction of SNP chip technology, coupled with the success of the International Haplotype Map (HapMap) Project (<http://hapmap.ncbi.nlm.nih.gov>), has led to an explosion of new research studies involving whole or partial genome-wide scans, commonly referred to as genome-wide association studies (GWAS). High-throughput SNP genotyping platforms, including Affymetrix and Illumina chips, provide for simultaneous genotyping of 50,000 to one million SNPs. While a mammalian genome consists of approximately 3×10^9 bases, variability in the genome may be captured by a subset of SNPs due to well-defined LD blocks and it has been proposed that whole genome-wide arrays of approximately one million SNPs are sufficient to characterize human genetic variability across a population. Several software tools have been developed that apply these methods while accounting for the computational demands of whole genome-wide investigations: such as PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/>, [54]), SNPAssoc (<http://www.creal.cat/jrgonzalez/software.htm>, [55]) and GenABEL (<http://www.genabel.org>, [56],[57]), haplostats (<http://cran.r-project.org/web/packages/haplo.stats/index.html>). PLINK is a stand-alone software written in C++ (but it can handle in house written R plugins) while SNPAssoc, haplostats and GenABEL are R packages available via CRAN (The Comprehensive R Archive Network, <http://cran.r-project.org>) and the latter has been extensively used in this work. The snpMatrix package is a component of the BioConductor open-source and open development software project for the analysis of genomic data [58]. GNU R is a free software language and environment for statistical computing and graphics (www.r-project.org).

2.3.4.1 Genotyping errors

A genotyping error is defined as a deviation between the true underlying genotype and the genotype that is observed through the application of a sequencing approach. These errors occur with varying degrees of frequency across the different technological platforms and arise for a variety of different reasons [59]. The most common statistical approach to identifying genotyping errors in population-based studies of unrelated individuals is testing for a departure from HWE at each of the SNPs under investigation [60]. This proceeds using either an asymptotic χ^2 -test or Fisher's exact test for association. There are a few notable drawbacks to this approach that we describe here, as

they are important to keep in mind in making the decision to remove SNPs from analysis: using tests of HWE exclusively for identifying genotyping errors is that a departure from HWE may in fact be a result of population substructure. Removing these SNPs from the analysis can therefore lead to a loss of data that are potentially informative regarding underlying structure in our population and thus data preprocessing may become too stringent. Another drawback comes in disease studies deviations from HWE within individuals with disease (cases) can be due to association between genotypes and disease status for disease studies. In order to account for this possibility, some researchers advocate testing for departures from HWE only within the control population. This, however, can be misleading as well since removing cases from the analysis may lead to an apparent departure from HWE when indeed the entire population is in HWE. One resolution to this problem for case-control studies involves application of a goodness-of-fit test to identify the most probable genetic disease model; e.g., dominant, recessive, additive or multiplicative [61]. This approach provides a means of distinguishing between departures from HWE that are due to the model for disease prevalence and those departures that are indeed a result of other phenomena, such as genotyping errors. Finally, multiplicity is also a challenge in this setting: multiple testing leads to an inflation of the type-1 error rate. Therefore, if we test for a departure of HWE at each of multiple SNPs, then the likelihood of incorrectly rejecting the null hypothesis (and concluding HWD) can be substantial, particularly in the context of GWAS. This is attenuated to some extent by the correlated nature of the tests, arising from LD among SNPs; however, consideration of multiple testing adjustments is still warranted. In this setting, extreme thresholds for significance, such as 0.005 or 0.001, are common. If a departure from HWE is detected then the entire SNP is typically removed from analysis: identification of specific records (individuals) within which an error exists are not provided by these approaches. Instead, it is assumed that the genotype information is incorrect across all individuals in the sample and the entire corresponding column is removed prior to statistical analysis.

2.3.4.2 Estimation of kinship matrix from dense SNP data

Kinship can be estimated from pedigree data using standard methods, for example the kinship for two outbred sibs is $1/4$, for grandchild-grandparent is $1/8$, etc. For an outbred person, the kinship coefficient is $1/2$ – that is two gametes sampled from this person at random are IBD only if the same gamete is sampled. However, in many situ-

ations, pedigree information may be absent, incomplete, or not reliable. Moreover, the estimates obtained using pedigree data reflect the expectation of kinship, while the true realization of kinship may vary around this expectation. In presence of genomic data it may therefore be desirable to estimate the kinship coefficient from these, and not from pedigree. It can be demonstrated that unbiased and positive semi-definite estimator of the kinship matrix [62] can be obtained by computing the kinship coefficients between individuals i and j with

$$\hat{f}_{ij} = \frac{1}{L} \sum_{l=1}^L \frac{(g_{i,l} - p_l)(g_{j,l} - p_l)}{p_l(1 - p_l)} \quad (7)$$

where L is the number of loci, p_l is the allelic frequency at l -th locus, $g_{i,l}$ is the genotype of j -th person at the l -th locus, coded as 0, 1/2, and 1, corresponding to the homozygous, heterozygous, and other type of homozygous genotype ([57],[62],[63]) The frequency is computed for the allele which, when homozygous, corresponds to the genotype coded as "1".

2.3.4.3 Analysis of structured populations

Populations may be structured in at least two ways. First, subjects may belong to rather different genetic populations (e. g. humans from China and Africa) or they may be sampled from a family in genetically isolated population. In the first case we need to consider that mean and variance of a trait under investigation may have different distributions (due maybe to environmental causes) while in the family based study environmental influences are more or less uniform for all study participants but they are characterized by high and variable kinship between each other. These two types of stratification may be expressed in terms of kinship: when we talk about different "populations", we assume very low kinship and high genetic differentiation between the members of these; when we talk about "relatedness", we assume high kinship and small degree of differentiation. This is illustrated at Figure 19: if we expect that study participants share common ancestors few (1-4) generation ago, the kinship between study participants is high, and the study can be classified as a family-based one (case); on the other hand, when common ancestors between study participants more remote the kinship between study participants is low, and such study may be classified as a population-based sample of "independent" people (cases 2 and 3).

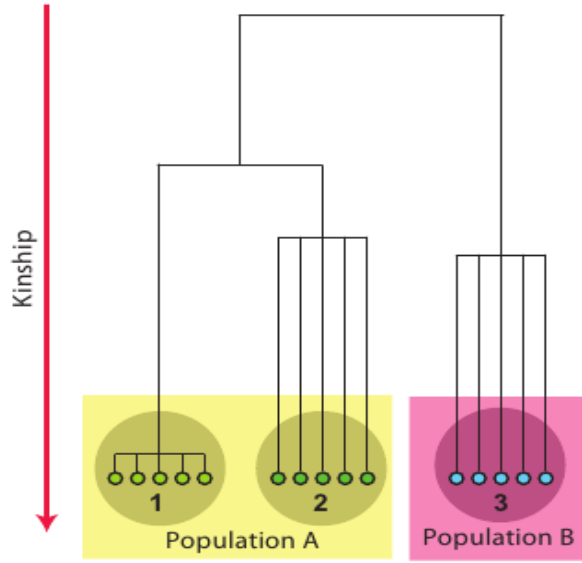


Figure 19: Structured populations

Figure 6.5: Three samples from two populations. 1: Family-based sample; 2,3: random sample of "independent" people

2.3.4.4 Genomic control

It follows that in presence of genetic structure standard association statistics may be inflated and the distribution is described as $\lambda \cdot \chi^2$; λ depends on the genetic structure of the sample, as characterized by pairwise kinship, and sample size. For binary trait analysis, it also depends on the composition of the case-control sample, as expressed by the proportion of cases/controls. For quantitative trait, it depends of heritability of the trait, and environmentally determined differences co-occurring with the difference in kinship. However, λ does not depend on the allele frequency. Therefore, for any particular study sample, if F_{ST} is constant over the genome, λ is also a constant. Therefore λ can be estimated from the genomic data, using "null" loci – a set of random markers, which are believed not to be associated with the trait. This estimate can then be used to correct the values of the test statistic at the tested loci – a procedure, known as "genomic control". The test statistics computed from these loci thus estimates the distribution of the test statistic under the null hypothesis of no association. Let us consider M "null" markers, and denote the test statistic obtained from i -th marker as T_i^2 . Given genetic structure determines the inflation factor $\lambda : T^2 \sim \lambda \cdot \chi^2$. The mean of a random variable coming from χ_1^2 is equal to 1; if a random variable comes from $\lambda \cdot \chi^2$, the mean would be λ . Thus we can estimate λ as the mean of the obtained "null" tests.

In practice, it is recommended to use the ratio between the observed median and the one expected for χ_1^2 :

$$\hat{\lambda} = \text{Median} \frac{(T_i^2)}{0.4549} \quad (8)$$

For the tested markers, the corrected value of the test statistic is obtained by simple division of the original test statistic value on λ : $T_{corrected}^2 = T_{original}^2 / \lambda$. The genomic control procedure is computationally extremely simple – one needs to compute the test statistic using a simple test (e.g. score test), compute the median to estimate λ , and divide the original test statistics values onto λ . How to choose "null loci" is a GWA study? In a genome, we expect that a small proportion of markers is truly associated with the trait. Therefore in practice, all loci are used to estimate λ . Of cause, if very strong (or multiple weak) true associations are present, true association will increase the average value of the test, and genomic control correction will be conservative.

2.3.4.5 Analysis of family data

For pedigree-based data coming from genetically homogeneous population it can be shown that λ is a function of trait's heritability and pedigree structure, expressed as kinship matrix. Thus, genomic control is a simple and valid method to study association in genetically homogeneous families. However, this method reduces all the abundant information about heritability and relationship into a single parameter λ , therefore it is not the most powerful method. In quantitative genetics, a mixed polygenic model of inheritance may be considered as a valid alternative since it has sound theoretical bases and describes well inheritance of complex quantitative traits. It is assumed that for each individual its "individual" random residual e_i is distributed as Normal with mean zero and variance σ_e . As these residuals are independent between pedigree members, the joint distribution of residuals in the pedigree can be modelled using multivariate normal distribution with variance-covariance matrix proportional to the identity matrix : $e \sim MVN(0, I \sigma_e)$. The model may expressed as:

$$Y = \mu + G + e \quad (9)$$

trait values are equal to the sum of grand population mean, contribution from multiple additively acting genes of small effect G , and residual error e_i . Under this assumption, the distribution of the trait in a pedigree is described by a multivariate normal distribu-

tion with number of dimensions equal to the number of phenotyped people, with the expectation of the trait value for some individual i equal to $E[y_i] = \mu$ where μ is grand population mean (intercept). The variance-covariance matrix is defined through its elements V_{ij} – covariance between the phenotypes of subjects i and j :

$$V_{ij} = \begin{cases} \sigma_e^2 + \sigma_G^2 & \text{if } i = j \\ 2 \cdot f_{ij} \cdot \sigma_G^2 & \text{if } i \neq j \end{cases} \quad (10)$$

where σ_G^2 is variance due genes, f_{ij} is kinship between persons i and j , and σ_e is the residual variance. The proportion of variance explainable by the additive genetic effects, is termed (narrow-sense) "heritability", h^2 :

$$h^2 = \frac{\sigma_G^2}{\sigma^2} \quad (11)$$

Fixed effects of some factor, e.g. SNP, may be included into the model by modifying the expression for the expectation, e.g. :

$$E[y_i] = \mu + \beta_G g \quad (12)$$

leading to so-called "measured genotypes" model [64]. In essence, this is a linear mixed effects model – the one containing both fixed (e.g. SNP) and random (polygenic) effects. A large body of literature is dedicated to this class of models, and finding the solution for such regression represents no great methodological challenge. Applicability of this model to GWA analysis was first proposed by Yu and colleagues [65].

2.3.4.6 The FASTA method

The polygenic component G in the mixed model approach describes the contribution from multiple independently segregating genes all having a small additive effect onto the trait (infinitesimal model). For a subject for whom parents are unknown, it is assumed that G_i is distributed as Normal with mean zero and variance σ_G . Assuming a model of infinitely large number of genes, it can be shown that given polygenic values for parents, the distribution of polygenic effects in offspring follows Normal distribution with mean $(G_m + G_f)/2$ and variance $\sigma_G/2$, where G_m and G_f are the maternal paternal polygenic values. From this, it can be shown that jointly the distribution of polygenic component in a pedigree can be described as multivariate normal with variance-covariance matrix proportional to the relationship matrix Φ :

$G \sim MVN(0, \Phi \sigma_G)$. Thus the log-likelihood for this model can be written as a function of three parameters:

$$L(\mu, \sigma_G^2, \sigma_e^2) = \frac{-1}{2} \ln(|(\Phi \sigma_G^2 + I \sigma_e^2)|) + (Y - \mu)^T \cdot (\Phi \sigma_G^2 + I \sigma_e^2)^{-1} \cdot (Y - \mu) \quad (13)$$

where μ is the intercept, σ_G is the proportion of variance explained by the polygenic component, and σ_e is the residual variance. Covariates such as sex, age, or a genetic marker studied for association can be easily included into the model:

$$Y = \mu + G + \sum_j^n \beta_j C_j + e \quad (14)$$

where, C_j is the vector containing j -th covariate and β_j is the coefficient of regression of Y onto that covariate. This mixed model leads to the following likelihood expression:

$$L(\mu, \sigma_G^2, \sigma_e^2, \beta_1, \dots, \beta_j, \dots, \beta_n) = \frac{-1}{2} \ln(|(\Phi \sigma_G^2 + I \sigma_e^2)|) + (Y - (\mu + \sum_j^n \beta_j C_j))^T \cdot (\Phi \sigma_G^2 + I \sigma_e^2)^{-1} \cdot (Y - (\mu + \sum_j^n \beta_j C_j)) \quad (15)$$

For example the effect of an allele in a SNP could be included as:

$$Y = \mu + G + \beta_g g + e \quad (16)$$

where g is the vector containing genotypic values. In this mode, you can specify a variety of 1 d.f. models by different coding of the vector g . For example, with an AG polymorphism estimation of the additive effect of the G allele, should be done coding AA as 0, AG as 1 and GG as 2. Under this coding, the β_g will estimate additive contribution from the "G" allele. On the other hand a dominant model for G is desired, genotype coding should be AA and AG as 0 and GG as 1. A general 2 d.f. model, the specification such as:

$$Y = \mu + G + \beta_1 I_{(g=1)} + \beta_2 I_{(g=2)} + e \quad (17)$$

would estimate trait's deviation in these with $g=1$ (β_1) and these with $g=2$ (β_2) from the reference ($g = 0$). The classical way to estimate mixed polygenic model and test for significance are Maximum Likelihood (ML) or Restricted ML (REML) using equation (7.3). However, when large pedigrees are analysed, ML/REML solution may take prohibitively long time, i.e. from minutes to hours for single SNP analysis, making study of hundreds of thousand of SNPs impossible. Therefore fast approximate tests were developed for the purposes of GWA association analysis in samples of relatives. One of these is the FAmily-based Score Test for Association (FASTA, [66]). It is based

on the classical polygenic mixed model and performed in two steps. First, polygenic model as specified by equation (7.1) and likelihood (7.2) is estimated using available data. Secondly, the maximum likelihood estimates (*MLEs*) of the intercept, μ , proportion of variance explained by the polygenic component, σ_G , and residual variance, σ_e , are used to compute the FASTA test statistics:

$$T_F^2 = \frac{((g - E[g])^T \cdot (\Phi \sigma_G^2 + I \sigma_e^2)^{-1} \cdot (Y - \mu))^2}{((g - E[g])^T \cdot (\Phi \sigma_G^2 + I \sigma_e^2)^{-1} \cdot (g - E[g])^T)} \quad (18)$$

It can be shown that T_F follows χ^2 when pedigree structure is 100% complete and 100% correct. As this is never actually the case, application of GC to correct for residual inflation is recommended. FASTA test results in unbiased estimates of the SNP effect and correct P – values. However this is correct only when alternative is reasonably close to the null, i.e. when the SNP explains small proportion of trait's variance. Disadvantages of this test are that it can be relatively slow when thousands of individuals are analysed, and that permutation procedures can not be applied to estimate genome-wide significance, because the data structure is not exchangeable.

3 De novo assembly of water buffalo genome

3.1 Introduction

Water buffalo (*Bubalus bubalis*) was domesticated 3000-6000 years ago, and since then it has been of economic importance as dairy, meat, and draught breed[67]. Although water buffaloes are terrestrial animals, they spend a good portion of time wallowing in mud holes or rivers in order to keep cool and they typically inhabits rivers, streams, or mud holes. Buffaloes are thought to have been introduced into Italy from central Europe in the sixth century or by the Bay of Tunis in the seventh century at the same time as the Arab conquests. Introduction of water buffalo populations into Australia, Africa, and the Americas has taken place only recently. Not only are they draught animals, but they are also a source of meat, horns, skin and particularly the rich and precious milk that may be converted to creams, butter, yogurt and many cheeses. There are about 170 million water buffalo heads (compared to 1.3 billion cattle) in the world with about 161 million in Asia, 3.7 million in Africa, 3.3 million in South America, and the rest are distributed in Europe and Australia. Water buffalo provide ~5% of world milk production.

3.2 The water buffalo data set

DNA sequencing for the water buffalo genome assembly was made using funding from the International Buffalo Genome Consortium and performed by Sonstengard et al. at the USDA-ARS laboratory (<http://www.ars.usda.gov/main/main.htm>) laboratory in Maryland. The subject chosen for sequencing was pure breed Mediterranean female buffalo (the fascinating Olympia, Figure 20) from a an herd in Italy and tissue sample were donated by the owner. The water buffalo genome dataset was a mixed Illumina/454 data set which has proven to be the best data type choice for costs and quality of the final assembly [68]. It was composed of three type of data:

1. Illumina paired end reads, which give the real genome coverage and are used to build the De Bruijn graph; these are 100 bp long reads with a 200 bp insert size; a total of 571,334,795*2 reads were present in the data set

3. De novo assembly of water buffalo genome

2. Illumina mate pair reads: 75 bp long reads obtained sequencing the ends of a large (up to 10 Kbp) DNA molecule; these are useful during contig and scaffold building but are affected by chimerism that causes the reads to be not always the true ends of the starting molecule; chimeric reads must be removed to be useful; a total of 167,677,444*2 reads was initially present in the data set with insert size between 4 and 6 Kbp
3. Roche 454 single end and mate pairs, 400 bp long and with insert size up to 30 Kbp; used in the same way as Illumina mate pairs but not affected by chimerism; a total of 10,228,343 unmated and 2,416,466*2 mated reads were present both 454 and Illumina reads are used only in the last part of the assembly and too high coverage of this data maybe detrimental and certainly useless; for this reason they are cleaned from redundancy before the assembly.

Initial clone coverage before any data manipulation was about 40X for Illumina short paired ends and 1/2X for Illumina mate pair libraries (jump libraries hereafter) and Roche 454. As already explained in section 2.1.3.4 reads with long insert size is used for jumping over repeats and for joining scaffolds and high coverage of this kind of data is useless and may be detrimental to the quality of the assembly (in addition to being very costly). Actual coverage of the buffalo genome is obtained with paired ends.



Figure 20: fascinating Olimpia

3.3 Roche 454 preprocessing

The first step performed has been the cleaning of Roche 454 data, including mate pairs and single end libraries. The procedure for 454 is shown in Figure 34:

1. 454 reads which came in the native Roche .sff format were converted into the CABOG fragment file (.frg) format using the *sffToCA* script that is shipped with the Celera assembler. This tool check for proper read length and base calls quality and also for the presence of the linker when dealing with mate pairs. Converted reads are trimmed and proper mate pairs are created in the output .frg files.
2. Using *only* 454 data a genome assembly is performed with CABOG up to the creation of unitigs; this involves further data correction and trimming performed by Celera assembler.
3. Finally, reads are mapped unto unitigs and then are removed if their alignment coordinates are the same. 454 data is unaffected by chimerism (see below) like Illumina mate pairs.

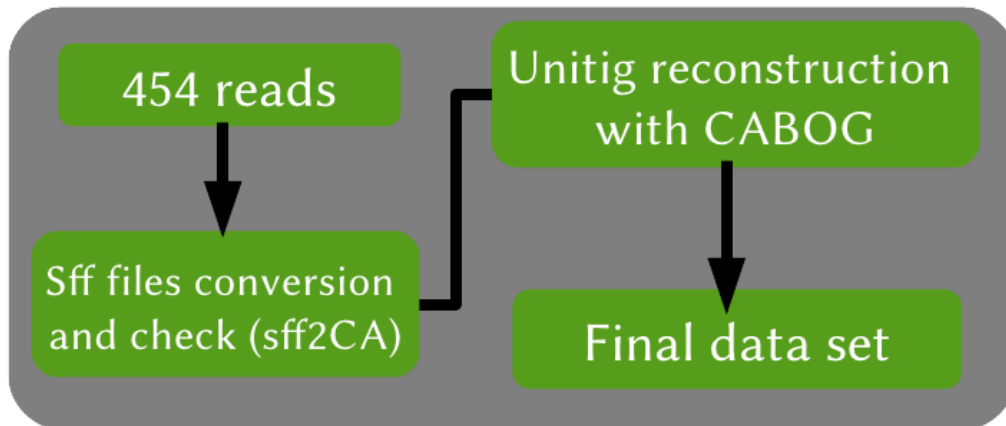


Figure 21: 454 preprocessing workflow

In this case data turned out to be of very high quality (see) since very few reads were discarded. Most of the cleaning was made in step 1 (.sff to .frg conversion) and only 1% of reads were found out to be chimeric. Remaining 454 reads were then collected in a unique fragment file and stored for the late steps of the MSR-CA procedure after the creation of unitigs with Super-reads.

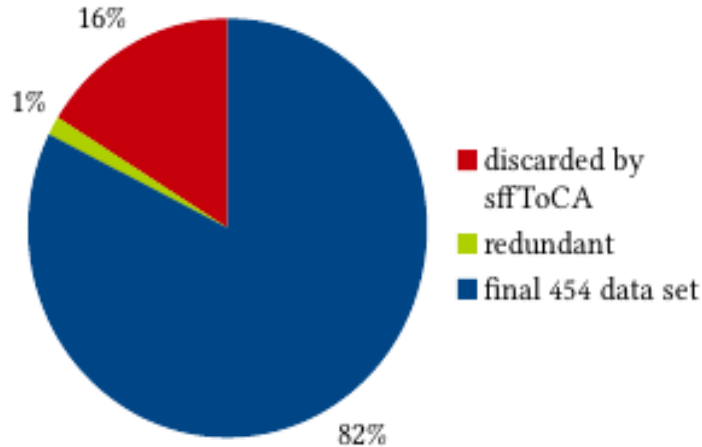


Figure 22: Roche 454 preprocessing results

3.4 Error correction and Super-reads creation

As already explained (see sections 2.1.2 and 2.1.3.3) reads may contain lower quality base calls in particular at their 3' ends. The problem is (partially) solved processing read prefixes more than their suffixes. This error-correction step is of fundamental importance for the construction of a reliable De Bruijn graph that contains as

few as possible errors and artificial edges. Prefixes and suffixes (i.e. reads 3' and 5' ends) can map to multiple paths. By a heuristic, mappings that are significantly better than their second-best choice are preferred. Just as the suffixes would add coverage to a multiple sequence alignment, they add connectivity to the graph. For this purpose we used a new k-mer counting algorithm and associated implementation, called Jellyfish, which is fast and memory efficient. It is based on a multithreaded, lock-free hash table optimized for counting k-mers up to 31 bases in length. Due to their flexibility, suffix arrays have been the data structure of choice for solving many string problems. For the task of k-mer counting, important in many biological applications, Jellyfish offers a much faster and more memory-efficient solution.

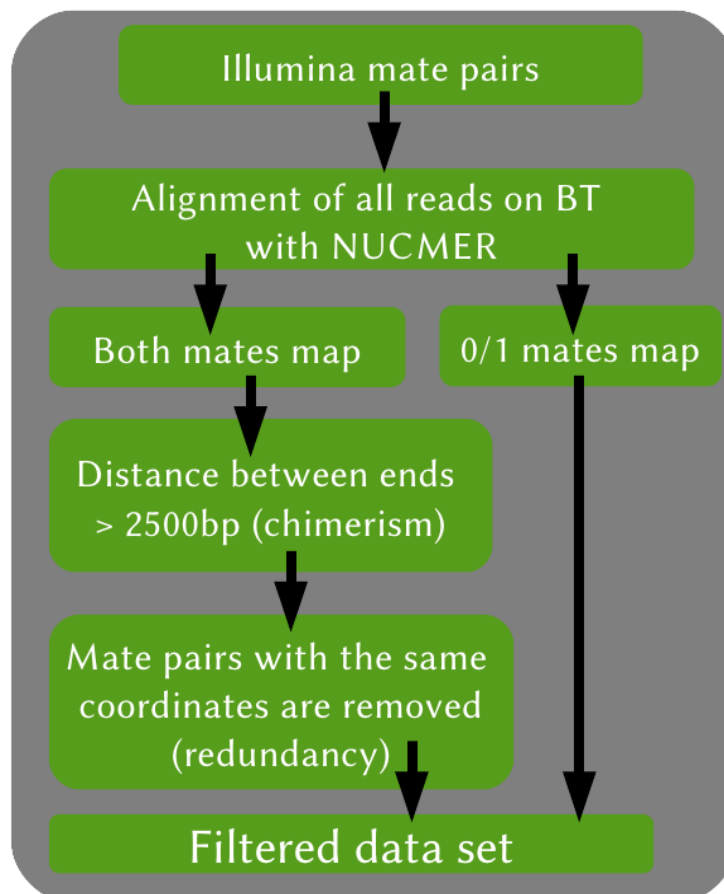


Figure 23: *Illumina jump libraries preprocessing workflow*

3.4.1 Illumina Mate Pair preprocessing

The following step was to clean Illumina mate pairs and chimerism had to be taken into account. Illumina mate pairs are obtained slicing DNA into 2-6 Kbp long frag-

ments which are then circularized again after attaching biotin labelled dNTPs. Non circularized molecules are removed by digestion and DNA is fragmented into small pieces. Proper Mate pairs are those reads located at the end of the biotin junction and oriented outwards (outties) so that their insert size is about the length of the circle; chimeras are mate pair reads that are sliced in a wrong point of the DNA molecule after circularisation (see Figure 24)

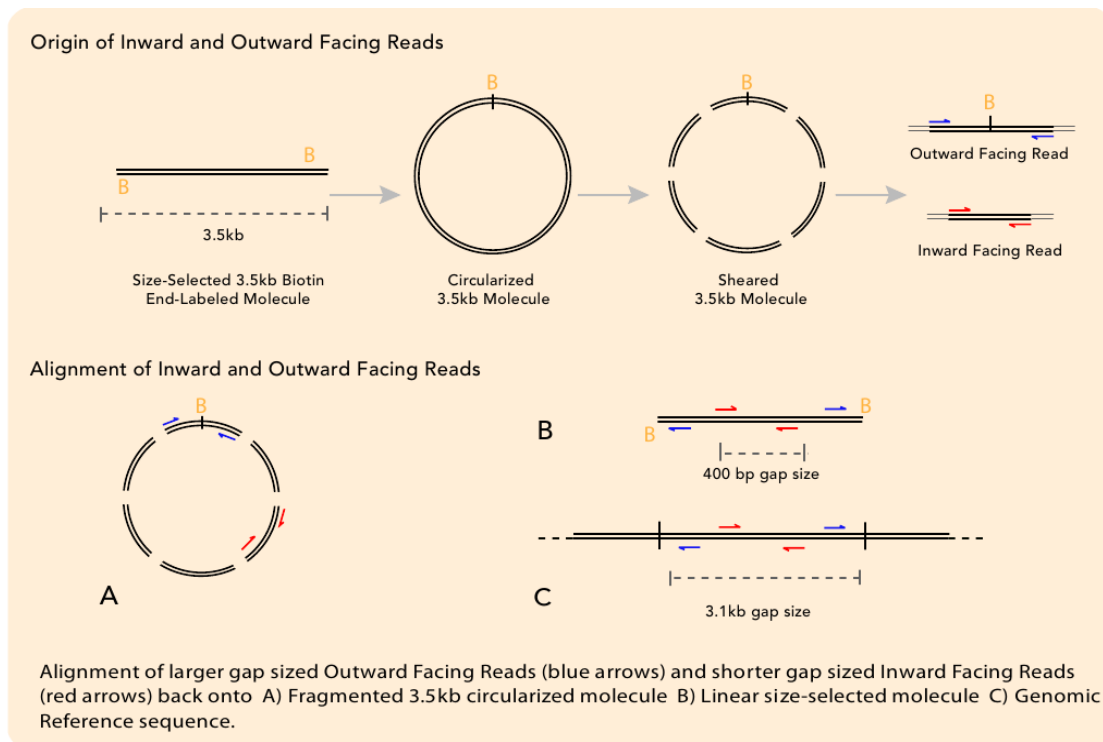


Figure 24: Origin and Alignment of Inward (innies) and Outward Facing (outties) reads in Mate Pair libraries

The filtering protocol is explained in Figure 23. Basically, mate pairs are mapped onto *Bos taurus* genome chromosomes with the NUCmer [69], [70] aligner and if both mates maps on the same chromosome the distance between the outside 5' and 3' end, i.e. the insert size, is calculated. Reference sequence was the latest cattle genome version assembled at University of Maryland, UMD 3.1 [68]. If the distance is too high the mates are considered non mapping otherwise they discarded because of chimerism if distance is less than 2.5 Kbp and because they are redundant is a mate pair with the same coordinates is present in the data set. The results of the alignment are shown in Figure 25 where the distance between mapping reads based on the coordinates of *Bos taurus* genome is depicted. The distribution of reads according to the insert size of lib-

raries is clearly visible; similarly the high peak at very distances (containing about a third of the initial reads) clearly points out the existence of chimeric reads. Final results yielded about 60% of mates which mapped with both ends and 30% of initial mate pairs that were conserved.

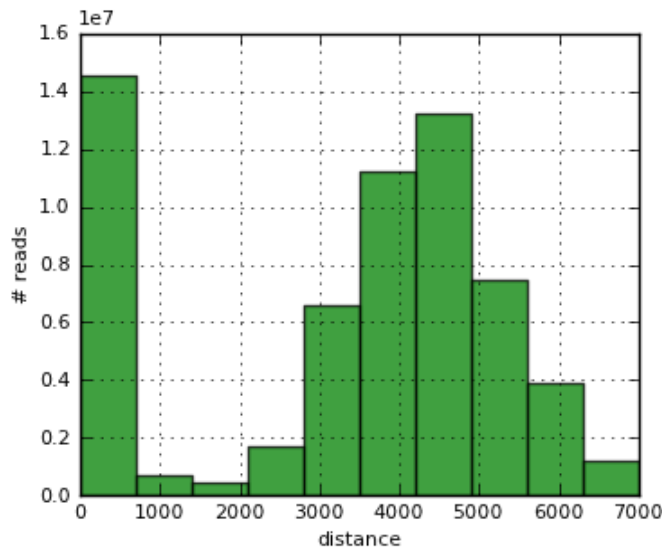


Figure 25: Alignment of Illumina jump libraries on *Bos taurus* genome

3.4.2 De Bruijn graph construction and error correction

After the 454 data preprocessing the Illumina short paired end libraries, which formed the bulk of available data were processed. This includes k-mer counting, performed with Jellyfish, and error-correction of base calls in reads based on the k-mer hash table. Illumina jump libraries too are error-corrected even if they are not used in the hash table generation. Error-correction involves base change, read trimming and complete read elimination if it is of very poor quality. Since mates are processed independently this results in a number of reads losing their mate; this is not a real issue with paired end since the goal is to merge them into Super-reads but a great deal of jump libraries (which are more error-prone) are discarded in this way since they are useful for the information that comes with the mating. After this step paired end and jump libraries are processed independently since the former are used to build Super-Reads and the latter are further refined using alignment over a similar species such as cattle.

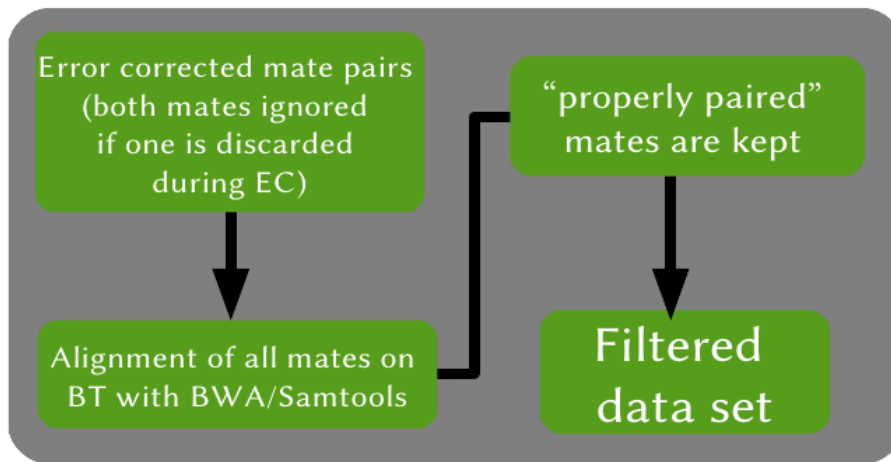


Figure 26: Alternative jump libraries preprocessing workflow

3.4.3 Jump libraries preprocessing after error-correction

Because a high number of reads did not map on the cow genome as expected with tested another method (Figure 26) consisting of doing error-correction before aligning the reads. This yields the following advantages: the number of reads is smaller because error-correction will drop some mates and we only consider both existing mates librar-

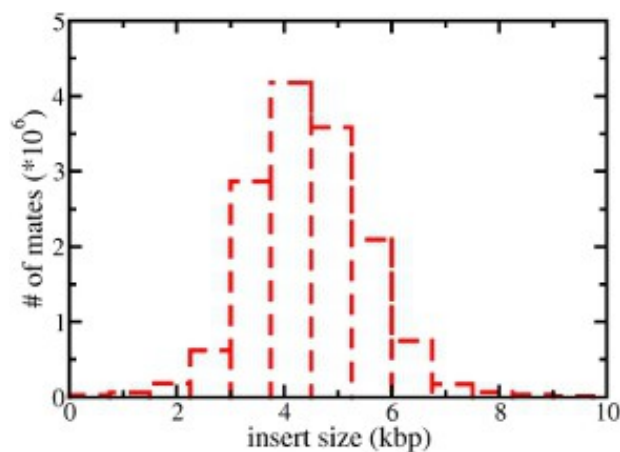


Figure 27: Results of mapping on *Bos taurus* after error correction- 1

ies for which both mates are present; mappings are of higher quality and more reads will map; in this way it is possible to include pairing information in the aligning, use different packages such as BWA (Burrows-Wheeler Aligner, [71], [72]) which are less flexible than NUCmer but generate cleaner data. The results of this alternative procedure are shown in Figure 27 and it is clear that once errors are removed all remaining reads (~20%) maps on *Bos taurus* and only a few are chimeric. The final jump library

data set contained 58,679,256 reads → 29,339,628 mate pairs → 17.5% of initial (un-corrected) data.

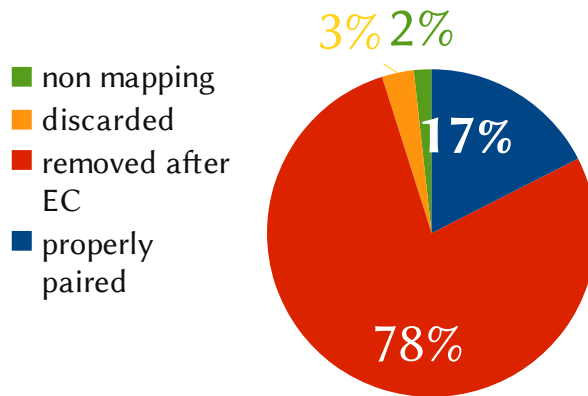


Figure 28: Results of mapping on *Bos taurus* after error correction - 2

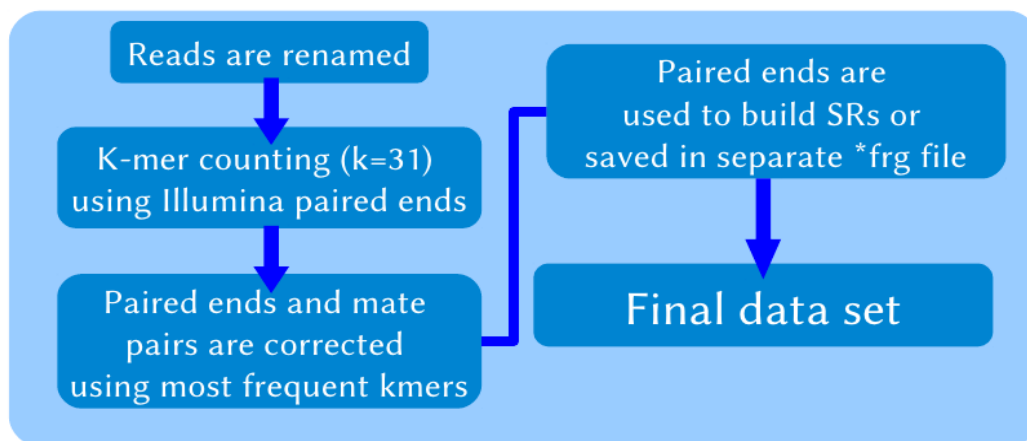


Figure 29: Super-reads creation workflow

3.4.4 Super-reads creation

Finally, paired end reads, according to the procedure described in Figure 29, were merged into Super-reads after the error-correction (in which a mere 2% of reads was removed); 63% of reads were merged in a total of ~40M Super-Reads yielding a data set size reduction of ~30X; the remaining paired end were used in the original form (Figure 30). The dimension of Super-reads compared to normal data before and after the quality trimming can be observed in Figure 31 where it is clear that each Super-read sums up the information contained in several starting paired ends. Once that

Super-reads have created the assembly can move on the third part of the MSR-CA pipeline (see Figure 11) i.e. the “real” genome assembly performed using the OLC approach (section 2.1.3.2).

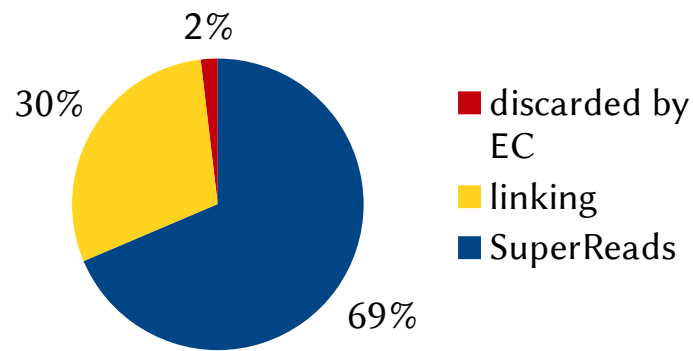


Figure 30: Super-reads creation results

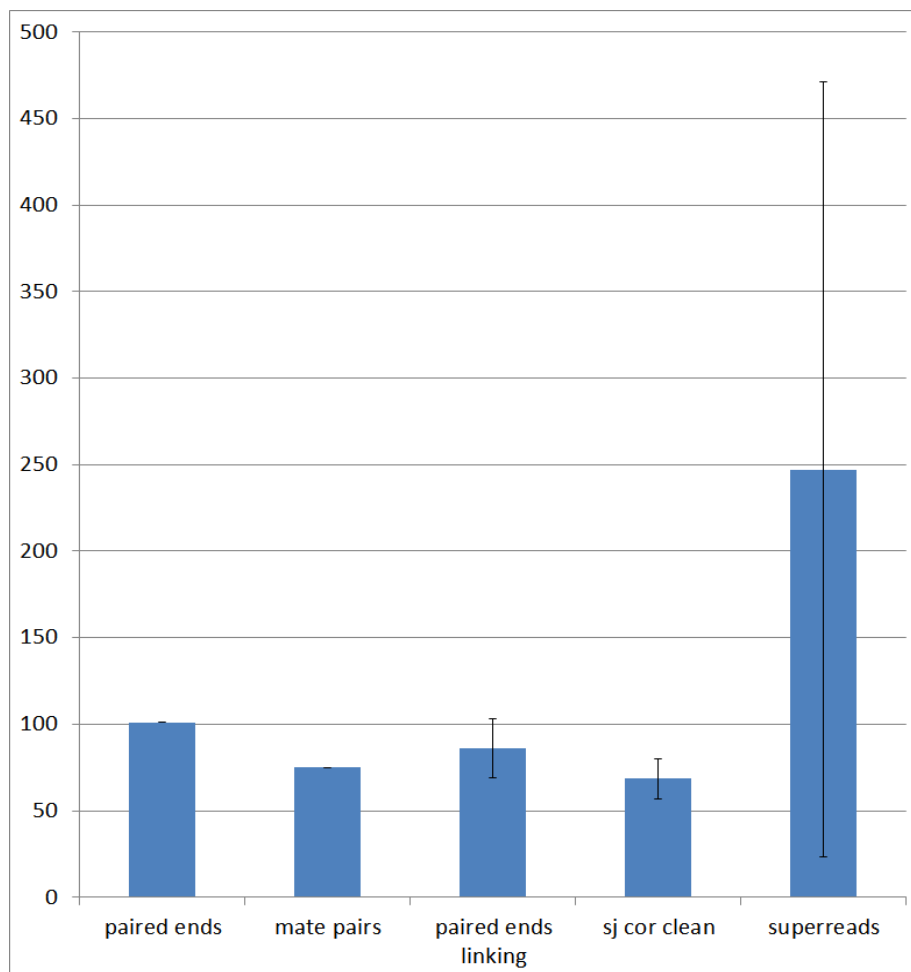


Figure 31: Super-reads dimension

3.5 Genome assembly using Super-reads

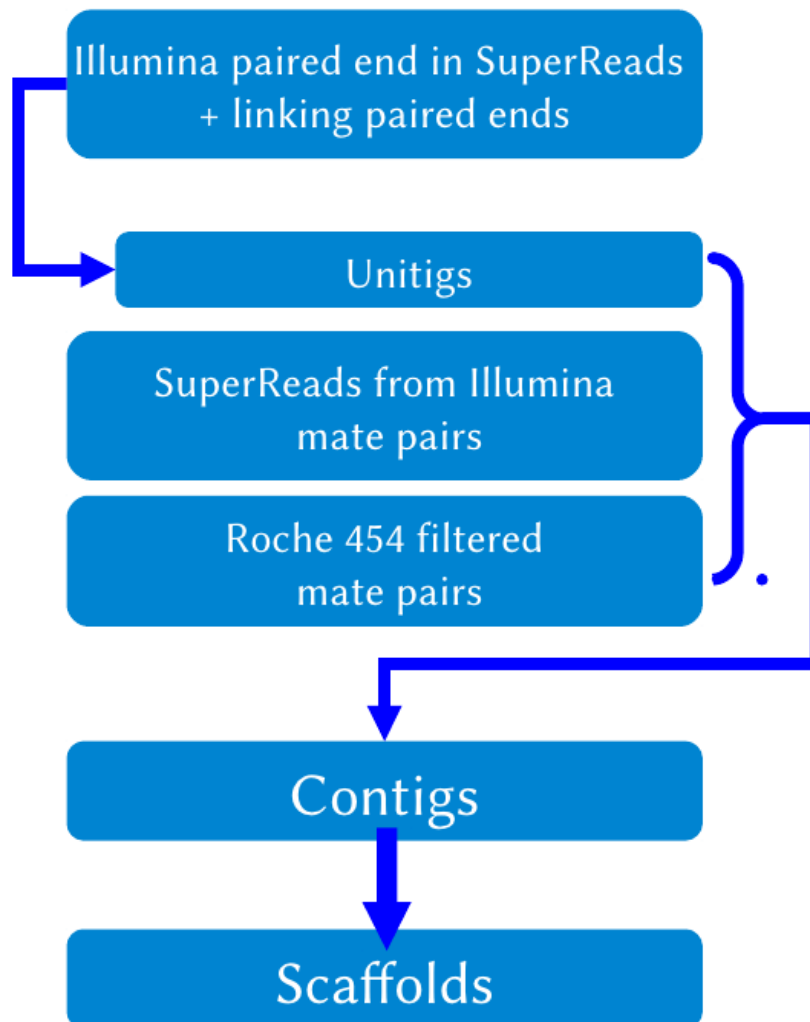


Figure 32: CABOG assembly workflow

This step is performed using the CABOG assembler. Its basic phases are shown in Figure 32. Since actual coverage of the genome is given by Illumina paired end libraries and thus by Super-reads (and the remaining linking paired end reads) after the first two steps only this kind of data is used to create unitigs. Then, contigs are created using reads with high insert size and then the CGW scaffolder is run. Calculations were initially started in Septe 2011 using MSR-CA version 1.1. Before running on actual buffalo the assembler was tested on a bacteria, *Rhodobacter sphaeroides*, comparing the obtained sequence with the reference genome. Meanwhile the real assembly was started and during k-mer counting (Jellyfish) several bugs were fixed in subsequent phases yielding MSR-CA version 1.2 and we performed also alignment of jump libraries on cattle genome. By the time we realized that the first approach was unoptimal

still more improvements had been made in the assembler (version 1.3.3) and the first assembly started with version 1.1 was still running. Thus, we started a new assembly using the second method for cleaning jump libraries and also we re-executed the Super-reads creation step since that part of the code had been changed and optimized; k-mer counting results were simply recycled since Jellyfish code had not changed. Summarising there are currently two (semi-)independent assemblies of water buffalo that running which differ mostly for the jump libraries Super-reads refined data sets while 454 data is unchanged. The first assembly is now finalizing scaffold creation and should finish in a few weeks given enough computational resources are available while the second one is running the first part of the CABOG pipeline and thus unitigs have been created yet. Although this assembly has been started much later given software improvement and much cleaner data it should run much faster than the previous one; however a precise forecast of time needed is very difficult to make because subtle differences in data can have tremendous impact on assembly performance, particularly in scaffolding.

3.6 Guided assembly aligning on *Bos taurus* genome

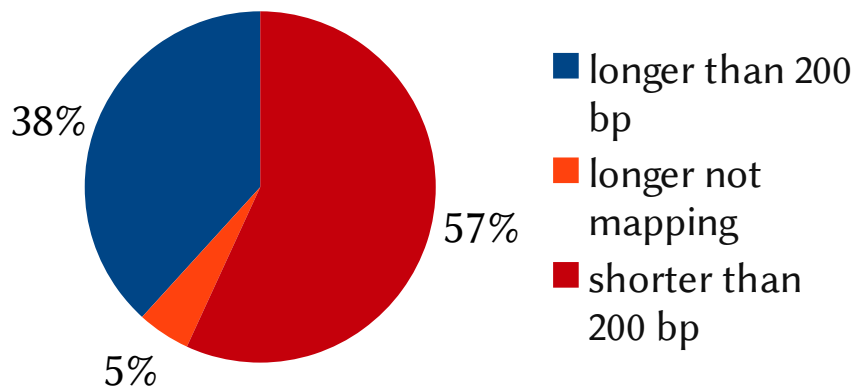


Figure 33: Guided assembly results

Cattle (*Bos taurus*) and water buffalo (*Bubalus bubalis*) are both members of the *Bovinae* subfamily. Thus, one expects their genomes to be fairly similar and one species may be used as model for the other. Starting on these premises an assembly of water buffalo aligning data on the cattle genome (guided assembly) was also performed

using long unitigs obtained in the first attempted *de novo* assembly of water buffalo even before the formation of the scaffolds, reconstructing the genomic sequence of each chromosome of the water buffalo, using *Bos taurus* genome (UMD 3.1, [68]) as a reference. Long unitigs (more than 200 bp) instead of starting data or Super-Reads were chosen for the following reasons:

1. Starting data contains errors that are subsequently removed after Error-correction and reads are shorter than Super-reads and unitigs above the chosen threshold
2. Super-reads do not include information from linking paired ends and may still contain error and information redundancy removed by CABOG

Thus, unitigs represent the smallest and cleanest set of read-like data that we could use to maximise quality and speed of the guided assembly. We started with 11,214,882 of unitigs Figure 33, 43% of which were longer than 200 base pairs. Most of these unitigs correctly aligned on cattle genome (using NUCmer); total length of cattle genome is ~ 2.63 Gbp and using only unitig data we covered 90% of this genome.

Scaffolds were created on each *Bos taurus* autosome (BTA) calculating consensus sequences of partially overlapping unitigs while unitigs completely contained in grater mappings (scaffolds or single unitigs) were discarded. Non covered zones were taken as gaps. Of course this procedure does not reflect that water buffalo has only 25 chromosome pairs and not 30 like cattle. In this way we obtained a draft guided assembly (still being completed) of water buffalo with a N50 contig size of 3182 (N50 is defined as the contig length such that using equal or longer contigs produces half the bases of the genome. The N50 size is computed by sorting all contigs from largest to smallest and by determining the minimum set of contigs whose sizes total 50% of the entire genome, <http://seqanswers.com/forums/showthread.php?t=2332>) which is very poor if compared to other *de novo* assemblies but may constitute a starting point for subsequent refinement of *de novo* assembled sequences; on the other hand these results are based on subset of total data (only long unitigs) and do not include 454 and jump libraries which are fundamental to improve contig and scaffold size.

3.7 Computational cost

The work presented here involves heavy computational cost and needs to be executed on large High Performance systems.

All steps of the MSR-CA protocol that involve data preprocessing out of the main pipeline i.e. 454 filtering and sequence alignment with NUCmer or BWA on *Bos taurus* genome were performed on the Matrix (<http://hpc.caspur.it/guides/the-hitchhikers-guide-to-the-matrix>) and Jazz (<http://hpc.caspur.it/guides/jazz-gpu-cluster>) Clusters present at CASPUR (<http://www.caspur.it/>). Cleaning of 454 took 10 days on a total of 16 nodes with 8 CPU and 32 GB of RAM each while alignment of reads on *Bos taurus* genome took 7 days on a total of 4 nodes with 8 CPU and 32 GB of RAM each (NUCmer) and 3 days on a single server with 12 CPUs and 48GB of physical memory. It is worth to observe that these results does not imply that BWA is faster than NUCmer since the latter was aligning a much greater number of reads one chromosome at time.

Actual MSR-CA calculations were performed on the Lagrange server made available at CILEA (<http://www.cilea.it/>) which has 32 physical CPUs (64 with Hyperthreading enabled) and 512 GB of RAM. Different versions of the software were tested on Lagrange running assemblies of *Rhodobacter sphaeroides* takes as a standard test (<http://gage.cbcb.umd.edu/index.html>). K-mer counting with Jellyfish was the most memory costly steps requiring up to 400G of allocated memory for approximately 3 days. Error-correction and Super-Reads creation took approximately 10 days running on 32 CPUs (Hyperthreading enabled) and allocating up to 256 GB of memory when using the version 1.2 of MSR-CA and about two weeks on node of the Jazz cluster with 96 GB of RAM with version 1.4.

As for the two assemblies calculations still running the first took approximately 40 days of elapsed time on a single CPU on Lagrange requiring between 50 and 100 GB of RAM during scaffolding.

Total storage required for the project includes 400 GB to store (compressed) sequence data while the assembly at the scaffolding step occupied more than 7 TB of disk space even with careful storage management.

4 Cattle quantitative genetics

4.1 The SELMOL data sample

4.1.1 Animals and phenotype data

The initial sample was formed by 2935 bulls of two dairy, two beef and one double purpose breeds: 761 Italian Brown, 899 Italian Holstein, 323 Piedmontese , 464 Marchigiana and 488 Italian Pezzata Rossa bulls, respectively. Genomic DNA was extracted from semen using the NucleoSpin Tissue kit (Macherey & Nagel) according to manufacturer's instruction. DNA was checked for quality on agarose gel and quantified using a DTX microplate reader (Beckman Coulter) after staining with Picogreen (Invitrogen). Samples were genotyped using BovineSNP50 Genotyping BeadChips (Illumina, San Diego, CA, USA). Phenotypic data was made available from each breed's national breeders association: Piedmontese Estimated Breeding Values (EBVs) for calving ease were provided by the Italian Piedmontese National Breeders Association (ANABORAPI, <http://www.anaborapi.it/>) while Italian Brown Daughter Yield Deviations for milk, protein and fat yield were provided by the Italian Brown national breeders association (www.anarb.it).

4.1.2 SNP panel and genotyping

The 50K SNP array contains 54001 SNPs distributed across the entire genome, with an average SNP spacing of 51 Kb and a proportion of known chromosome positions of about 97%; SNP positions within each chromosome were based on the *Bos taurus* genome assembly Btau_4.0 [73]. Genotyping was outsourced to Geneseek, USA (www.geneseek.com).

4.2 Candidate genes in Italian Brown breed

4.2.1 Introduction

Italian Brown is the second most popular dairy cattle breed in Italy, with a population of about half a million head and 160.000 cows registered in the national herd-book. Its

presence has been recorded in Italy since the XIV century. The breed was traditionally raised in small size herds in uplands and had originally a double beef/dairy attitude. Modern selection towards milk production, basically obtained importing Swiss Brown bulls from the U.S., resulted in increasing popularity and herd size. Italian Brown is characterized by a good balance between milk yield and quality and its milk is an ideal raw material for cheese production.

The genetic control of milk traits has often been described by an infinitesimal model, assuming the contribution of an infinitely large number of loci, each with a small effect. However, Quantitative Trait Loci (QTL) experiments conducted in recent years revealed that some genomic regions have sensible effects on milk-related traits in many cattle breeds. QTLs affecting milk production have been identified on BTA (Bos taurus autosome) 3, BTA6 and BTA14 using different methods ([74]). In this study, a functional candidate gene approach was performed to identify genes associated with milk production traits in 561 Italian Brown bulls [75].

4.2.2 Candidate gene selection

Candidate genes were chosen on the basis of evidence of physiological or biochemical processes related to milk production traits, or of significant association with production traits reported in previous association studies either in cattle ([42–50]) or in related species e.g., sheep and goats ([85]). A first panel of 106 SNPs was selected within these genes and in nearby chromosome regions, using either information available in the literature or *in silico*, after a database interrogation on NCBI dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>). SNP genotyping was outsourced to Kbiosciences (www.kbiosciences.co.uk).

4.2.3 Results

	Gene	Symbol	BT A	Position	#SNP
1	Acetyl-Coenzyme A Carboxylase α	ACACA	19	11	1
2	Butyrophilin, subfamily 1, member A1	BTN1A1	23	31	1
3	Corticotropin Releasing Hormone	CRH	14	31	2
4	α s1-Casein	CSN1S1	6	88	2
5	β -Casein	CSN2	6	88	1
6	κ -Casein	CSN3	6	88	1
7	Chemokine receptor 1	CXCR1	2	110	1
8	Fatty Acid Binding Protein 4	FABP4	14	41	1
9	Fatty Acid Synthase	FASN	19	52	2
10	Growth Hormone	GH	19	49	2
11	Growth Hormone Receptor	GHR	20	34	1
12	Leptin hormone	LEP	4	95	2
13	Leptin receptor	LEPR	3	85	1
14	β -Lactoglobulin	LGB	11	107	3
15	Milk Fat Globule-EGF factor 8 protein	MFGE8	21	20	1
16	Oxidized Low density Lipoprotein	OLR1	5	97	1
17	Prepro-Oxytocin	OXT	13	52	1
18	Peroxisome proliferator-activated receptor γ coactivator 1 α	PPAR- GC1A	6	44	3
19	Prolactin Receptor	PRLR	20	41	1
20	SRC v-src sarcoma viral oncogene homolog (avian)	SRC	13	67	1
21	Thyroglobulin	TG	14	7	1
22	Thyroid Hormone-Responsive Protein	THRSP	29	19	1

Table 1: Italian Brown candidate genes

Candidate genes retained after the filtering procedure and their acronyms (Symbol), the chromosome on which they are located (BTA), their position in millions of bases according to release 4.0 of NCBI bovine genome, and the number of informative SNPs analysed within each gene (#SNP).

4.2.3.1 Data Analysis

Subjects with less than 10% of missing genotypes were retained. SNPs quality control excluded markers showing more than 10% missing genotypes and minor allele frequency (MAF) lower than 2%. Hardy Weinberg Equilibrium (HWE) was assessed on filtered data set ($P < 0.01$). A total of 106 SNPs in candidate genes chosen on the basis of known relationships between physiological or biochemical processes and on evidence of their significant association with milk traits was genotyped on the whole population formed by 550 Italian Brown bulls. Of these markers, 49 resulted monomorphic in the analysed population; this left 57 polymorphisms located in 29 genes, as shown in Table 1. A total of 76 sires had a missing genotype rate over 10% and were eliminated. In addition, 4 SNPs having of number of missing genotypes $>10\%$ and 22 which had a minor allele frequency of less than 2% were discarded; no markers were found to be out of Hardy Weinberg Equilibrium using a P-value of 0.01. After the described filtering, 31 SNPs in 22 candidate genes located on 13 different chromosomes and geno-

typed on 473 sires were retained. Names of candidate genes retained, their approximate position in millions of bases according to the release 4.0 of the bovine genome of the National Center for Bioinformatics (NCBI), and the number of informative SNPs analysed within each gene are reported in Table 1. Additional information about each SNP is given in Table 2. Phenotypes was considered to be a function of SNPs genotypes (one marker at time) of polygenic effects modeled by relatedness of subjects and of the year of semen collection of the bull (grouped in three levels) since it spanned more than 30 years of continuous selection. Calculation were performed with GenABEL. Significance levels were corrected using the Benjamini-Hochberg method (False Discovery Rate; FDR) with a threshold of 0.1 [86]. Genotype distribution was considered significantly associated to phenotypes if $P \leq 0.05$. Haplotype frequencies for genes containing multiple markers were calculated using the haplo.stats R package. To test for association between phenotypes and SNP haplotypes, a linear model approach including only the (fixed) haplotype effect was used ([1]) using residuals from single SNP calculations to remove kinship and covariate effect. The most frequent haplotype for each gene was considered as a reference and less common haplotype effects were calculated against this reference. Haplotypes with a frequency less than 5% were included in a single variable.

4.2.3.2 SNP association to milk yield

Milk yield (Table 3) was significantly associated to four SNPs on three genes located on three different chromosomes: increased milk production was associated to the A allele of CNS1S1_26181, to the T allele of OXT_rs43114144, and to the C allele for both SNPs on the LGB gene. Genotypic frequencies were nearly comparable for OXT_rs43114144 while one homozygous genotype was completely absent in CNS1S1_26181. Stronger association was obtained for LGB_1740 ($P < 0.01$) and LGB_5261 ($P < 0.001$). All SNPs resulted significant after the correction for multiple testing. Increased milk production was associated to the C allele in both LGB markers, although genotypic distributions were somewhat different: frequencies were comparable in LGB_1740, whereas TT was less common in LGB_5261. The effect of the high yield AA genotype of CSN1S1 polymorphism was comparable to the general mean while a slightly greater increase of milk yield, was observed for the CC genotype in OXT_rs43114144. More relevant gains (about 25%) were associated to both LGB polymorphisms.

	SNP	BTA	Alleles	% missing	MAF	P-exact
1	ACACA_a	19	C/T	3.51	0.05	0.622
2	BTN1A1_b	23	G/C	2.89	0.13	0.31
3	CRH_a	14	G/C	3.30	0.46	0.71
4	CRH_b	14	G/A	2.68	0.02	0.20
5	CSN1S1_26181	6	A/C	3.30	0.08	0.06
6	CSN2_8101	6	A/G	2.89	0.27	0.56
7	CSN3_13068	6	C/A	6.39	0.31	0.74
8	CXCR1_rs41255712	2	C/T	3.51	0.11	0.63
9	FABP4_a	14	C/G	4.54	0.13	0.84
10	FASN_a	19	A/C	2.89	0.25	0.72
11	FASN_b	19	G/A	2.47	0.06	0.67
12	GH_5UTR_-303C/T	19	C/T	3.09	0.14	0.56
13	GH_e	19	A/C	3.92	0.03	1.00
14	GHR_b_AT	20	A/T	3.51	0.07	0.72
15	LEP_a	4	C/G	5.57	0.44	0.64
16	LEP_exon2_305C/T	4	C/T	2.89	0.40	0.34
17	LEPR_c	3	T/G	3.71	0.35	0.92
18	LGB_1740	11	C/G	2.89	0.41	0.57
19	LGB_5261	11	C/T	3.09	0.34	0.61
20	MFGE8_rs41974998	21	T/G	3.51	0.43	0.85
21	OLR1_a	5	C/T	2.89	0.09	0.41
22	OXT_rs43114144	13	C/T	2.68	0.46	0.20
23	PPARGC1A_a	6	A/C	2.68	0.22	0.89
24	PPARGC1A_rs43464407	6	A/C	5.98	0.03	1.00
25	PPARGC1A_c	6	G/A	2.68	0.08	0.75
26	PRLR_rs42321611	20	G/A	4.33	0.16	1.00
27	SRC_d	13	T/C	3.51	0.13	0.23
28	TG_rs41580467	14	G/A	3.30	0.37	0.55
29	THRSP_rs42321611	29	C/T	2.89	0.25	1.00

Table 2: Italian Brown SNPs statistics

The marker name, the chromosome on which it is located, the percentage of missing genotypes, the minimum allele frequency (MAF), and the out of HWE P-values are shown

SNPs	Genotype	Genotype frequency (%)	Genotype averages	P _{unadj}	Q _{FDR}
LGB_5261	C:C	42.68	0.8073	2.57E-03	7.48E-02*
	C:T	46.28	0.5572		
	T:T	11.04	0.3287		
OXT_rs43114144	C:C	27.54	0.5318	5.89E-03	7.96E-02*
	C:T	52.75	0.6609		
	T:T	19.70	0.7480		
LGB_1740	C:C	35.47	0.8552	1.05E-02	8.00E-02*
	C:G	47.01	0.5756		
	G:G	49.40	0.4404		
CSN1S1_26181	A:A	83.37	0.6664	1.10E-02	8.00E-02*
	A:G	16.63	0.5301		
Population average			0.6446		

Table 3: SNPs associated to milk yield

Genotype frequencies for each marker, average DYD value for each genotype (Genotype averages), unadjusted P-values (P_{unadj}) and multiple-testing corrected significance levels (Q_{FDR}) are displayed. The last row shows the average trait value calculated in the whole population.

* = SNP is significant after FDR correction

4.2.3.3 SNP association to fat yield

Fat yield was associated to four different SNPs (Table 4). Only one marker had a P-value lower than 0.01 and none of them was retained after FDR correction. The lowest significance level was shown by OXT_rs43114144; increased fat production was associated to the T allele (same in milk yield). Increased yield was associated to T in THRSP_rs42714483, and to the C allele of LGB_5261 and CSN3_13068; genotype frequencies were comparable for these two SNPs. Phenotypic averages calculated on production genotypes were comparable to the whole sample average with the exception of THRSP_rs42714483: in this case TT homozygous sires show high fat yield (increase of ~35% above the phenotypic population mean for fat yield) but the genotype is rare (less than 7% of sires) and the trait/genotype association failed to pass the multiple-testing correction.

SNPs	Genotype	Genotype frequency (%)	Genotype averages	P _{unadj}	Q _{FDR}
OXT_rs43114144	C:C	27.54	0.0196	3.67E-03	1.06E-01
	C:T	52.75	0.0246		
	T:T	19.70	0.0277		
THRSP_rs42714483	C:C	56.48	0.0201	2.19E-02	2.47E-01
	C:T	37.37	0.0273		
	T:T	6.16	0.0393		
LGB_5261	C:C	42.68	0.0279	3.90E-02	2.47E-01
	C:T	46.28	0.0222		
	T:T	11.04	0.0144		
CSN3_13068	C:C	47.58	0.0280	2.56E-02	2.83E-01
	C:T	43.39	0.0246		
	T:T	9.03	0.0051		
Population average			0.0240		

Table 4: SNPs associated to fat yield

4.2.3.4 SNP association to protein yield

Protein yield was significantly associated to LGB_1740, LGB_5261, CNS1S1_26181 and FASN_a polymorphisms (Table 5). Significance levels of the two LGB markers were very low ($P \sim 10^{-5}$), representing the highest allele/trait association obtained and only these two SNPs were retained by multiple-testing corrections. High protein yield was again associated to the C allele in both SNPs as for milk yield. The AA genotype in the CNS1S1_26181 marker was associated to increased protein yield. In FASN_a, increased protein yield was associated to the A allele and the AA genotype was also the most common: less productive CC individuals were less than 7% of the population.

SNPs	Genotype	Genotype frequency (%)	Genotype averages	P_{unadj}	Q_{FDR}
LGB_1740	C:C	35.47	0.0400	3.74E-005	1.08E-03*
	C:G	47.01	0.0277		
	G:G	49.40	0.0163		
LGB_5261	C:C	42.68	0.0422	7.92E-05	1.15E-03*
	C:T	46.28	0.0286		
	T:T	11.04	0.0216		
CSN1S1_26181	A:A	83.37	0.0331	9.81E-03	9.48E-02*
	A:G	16.63	0.0267		
FASN_a	A:A	55.84	0.0356	4.03E-02	2.92E-01
	A:C	37.37	0.0281		
	C:C	6.79	0.0234		
Population average			0.0319		

Table 5: SNPs associated to protein yield

4.2.3.5 SNP Haplotypes

The effect of marker haplotypes was estimated for four genes, retaining more than one polymorphism after the filtering procedure (see Table 1 and Table 2): Corticotropin Releasing Hormone (CRH, 2 SNPs), Fatty Acid Synthase (FASN, 2 SNPs), β -Lactoglobulin (LGB, 2 SNPs) and Peroxisome proliferator-activated receptor γ co-activator 1 α (PPARGC1A, 3 SNPs) genes. Linkage disequilibrium calculations showed very low inter- and intra- gene correlation between SNP alleles; only of the two β -Lactoglobulin polymorphisms had high r^2 value (0.817); inspection of haplotype frequencies (see below, SNP haplotypes subsection) confirmed that C alleles on the LGB_1740 and LGB_5261 SNPs were usually in phase. Of these markers only SNPs within the LGB gene were found to be significant in trait/single SNP associations. Both milk and protein yield were significantly associated to the TT haplotype of the two markers within the LGB gene; however the estimated effect was very small for

protein yield. Surprisingly the CA haplotype formed by FASN_a and FASN_b was associated to milk yield but not to fat yield.

SNP1	SNP2	Haplo-type	fre-quency	Trait	P	β
LGB_1740	LGB_5261	GT	0.3430	Milk yield	2.75E-02	-1.68E-02
LGB_1740	LGB_5261	GT	0.3430	Protein yield	1.43E-03	-5.47E-09
FASN_a	FASN_b	CA	0.0592	Milk yield	4.72E-02	2.95E-02

Table 6: SNP haplotypes significantly associated with milk or protein yield

SNPs in each haplotype, haplotype, haplotype frequency, trait for which the haplotype is significant and regression coefficients are shown.

4.2.4 Discussion

In this study we investigated the effects of SNPs within candidate genes on milk production traits in Italian Brown breed. Seven (out of 29) SNPs located within 6 different genes were found significantly associated with the examined traits (milk, fat and protein yield). The most significant results were obtained for markers within the β -lactoglobulin gene (LGB, located on BTA11), which is also the only gene showing more than one SNP significantly associated to all traits. Other significant SNP/trait associations were found within the CSN3 and CSN1S1 (both on BTA6), OXT (BTA13), FASN (BTA19) and THRSP (BTA29) genes.

SNPs were less strongly associated with fat yield than to the other two traits. Tables 3, 4 and 5 illustrate that in all cases extreme trait values were associated with a homozygous genotype, and heterozygous individuals show an intermediate behaviour (with the exception of CNS1S1_26181 where GG homozygous sires were absent), thus justifying the choice of an allelic model for the analyses. Two SNPs in the β -lactoglobulin gene, significantly associated to all studied traits, were in Linkage Disequilibrium and both showed increased milk and protein yield associated to the presence of the C allele. This was confirmed by the haplotype analysis in which a negative milk and protein yield of TT homozygote individuals was observed; however the estimated coefficients of the TT haplotype were very small in both cases, and the effect of the C allele in the two markers can be considered additive. Particularly important is their association with an increase of protein production (over 10%). β -lactoglobulin is the major whey protein of ruminant species and is able to bind a variety of ligands ([87]) such as cholesterol, but its biological role has still to be properly understood. Up to 11 LGB

variants and 50 polymorphisms [82] have been identified, and the presence of one variant or another has been related to its concentration in cow milk. β -lactoglobulin is of high nutritional value, and thus could represent a feasible target for selection. However, its concentration is negatively correlated to the concentration of caseins, representing a possible problem for the dairy industry when milk is finalized to cheese production. The OXT_rs43114144 T allele was associated to increased milk yield and its allelic frequencies are suitable for selection; this marker is located on the Oxytocin prepropeptide gene (OXT). Oxytocin is a nine amino acid long peptide, initially synthesized as a large molecular weight precursor which also consists of the oxytocin-carrier peptide neurophysin. Contraction of the mammary gland for milk ejection is one of main functions of OXT; The gene is necessary for successful lactation as well as for postpartum alveolar proliferation and mammary gland function [88]. The essential role of OXT for the milk let-down reflex has been demonstrated using OXT-deficient mice, which failed to nurse their offspring [89]. Increased milk and protein yield were also associated to the presence of the A allele in the CSN1S1_26181 polymorphism; however the observed allelic frequency of A was very high (92%) and thus there is small room for further selection. Although polymorphisms located within the Fatty Acid Synthase (FASN) were not found to be significant, the CA haplotype was slightly associated to increased milk yield. FASN is a multifunctional enzyme involved in the biosynthesis of long chain fatty acids with effects on fat-related traits in muscle and milk in cattle ([80],[90]) and polymorphisms in the FASN gene have been found to be related to fat depositions in milk and adipose tissue in beef and dairy cattle ([79],[83]) and sheep [46]. Unexpectedly, in our results this gene presents association with milk yield but not with fat yield [79]. However, the frequency of the CA haplotype is very low, making it unfeasible for selection.

4.3 Assessment of population structure in five Italian cattle breeds

4.3.1 Introduction

Present day cattle breeds are the result of years of human selection, adaptation to different environments and demographic effects as domestication, migration and selection, all contributing to the actual patterns of genetic diversity ([91–93]). During domestication process, breeds were selected for productivity traits as, for example, milk yield [94]. This anthropic selection has influenced the genetic structure of cattle

breeds, so that loci influencing potentially selected traits could be detected identifying regions of the genome under selection [75], [95]. Moreover, animal and semen exchange, carried out to improve production characteristics of breeds, have affected the genetics (and history) of the breeds [96]. The study of population structure and of relationships between cattle breeds can aid in reconstructing their history and can be relevant for conservation of local breeds. The majority of the studies on cattle genomic structure were focused on microsatellite and AFLP markers ([97–103]) Recently, the availability of high density SNP panels has given the possibility of analyzing the genetic structure of cattle populations with thousands of markers distributed across the entire bovine genome ([102], [104–107]).

The aim of this study was to analyze the genetic structure of five Italian cattle breeds using a 54K medium-density SNP panel. This was addressed by classical multi-dimensional scaling and by Bayesian inference methods. Moreover, we analysed the fixation index (F_{ST}) values in order to identify genomic regions potentially under selection. We considered a total of 2935 bulls belonging to two dairy (Italian Brown and Italian Holstein), two beef (Piedmontese and Marchigiana) and one double purpose, (Italian Pezzata Rossa) breeds. The Italian Holstein derives from Dutch and North America Holstein breeds imported in the late XIX century and is the most common dairy breed in the country. Piedmontese is a Northern Italy breed and it is as mix of several taurine breeds. Marchigiana is a breed from central Italy derived from Chianina and Maremmana breeds. The Italian Brown breed was originally a multi purpose breed reared in the Alps; it was selected from 1950 as a dairy breed by importing Swiss Brown bulls from the U.S. Pezzata Rossa is a beef/dairy breed imported from Swiss and herded mostly in North East Italy since the late XX century

4.3.2 Results

4.3.2.1 Preprocessing

Data was initially filtered using the GenABEL R package. Sires and markers with a call rate under 99% were discarded as well as SNPs having a minor allele frequency (MAF) lower than 5%. Sires were checked for abnormally high autosomal heterozygosity and discarded when showing a FDR $<1\%$ ([40], [108]). Then, sires of each breed were separated and Hardy Weinberg Equilibrium (HWE) was checked within each breed [109] setting a threshold of $P \leq 0.01$ in the filtered data set. Finally, data

was pooled again and the criteria for call rate, minor allele frequency and high autosomal heterozygosity were applied one more time.

Only autosomal SNPs and SNPs with complete map information were considered, leaving 51515 different markers out of the 54001 genotyped sires. A total of 138 sires and 12388 markers showing a call rate < 99% were excluded. Additional 8874 SNPs with a minimum allele frequency under 5% were discarded and seven sires were removed because they had too high autosomal heterozygosity. Then, a total of 1443 markers were discarded because they were out of HWE in at least one breed. No sires or markers were excluded after the second check on call rates and allele frequencies. The final complete dataset was thus formed by 2797 sires (755 Italian Brown, 861 Italian Holstein, 483 Italian Pezzata Rossa, 317 Piedmontese and 381 Marchigiana bulls) and 29848 SNPs.

4.3.2.2 Data analysis

Kinship among sires was estimated directly from genomic data as proposed by Astle and Balding [62]. Genetic distance between breeds and sires was estimated calculating a 0.5 – kinship between all sires and then applying classical multidimensional scaling to the complete matrix. STRUCTURE software v. 2.3.2.1 ([110], [111]) was used to analyse population structure. A total of 15,000 Markov chain Monte Carlo (MCMC) iterations (5,000 burn-in and 10,000 sampling) were performed for each tested K using the admixture model, considering allele frequencies correlated among populations and including no informative prior about individual membership; K values from 2 to 5 were used. Five independent runs for each tested K value were performed. The number of steps was chosen following Vaysse et al. [112] although a single run for each K with 50,000 steps to test the effect of longer runs was performed too. Graphical visualization of STRUCTURE results was performed by means of the DISTRUCT package [113]. To determine a genome-wide pattern of positive selection, the F_{ST} at each locus was calculated [52]. The loci under selection are expected to show an allele frequency that deviate from that of neutral loci, leading to an increased level of genetic differentiation. F_{ST} values were then plotted against genome location. Signatures of selection can be recognized when adjacent SNPs in a region show high F_{ST} [114], thus we used a sliding window approach, with a window of 8 SNPs. A region with high F_{ST} implies divergent selection between breeds, whereas low F_{ST} imply balancing selection between breeds. Fixation index was calculated using the method proposed by Nei and

Chesser [53] using in-house written R codes; this method was chosen because the sample includes (almost) all sires in the national herdbooks for three breeds (Marchigiana, Piemontese and Pezzata Rossa) and thus fixed effects errors of sampling seemed more important. Graphs were obtained using matplotlib (<http://matplotlib.sourceforge.net/>).

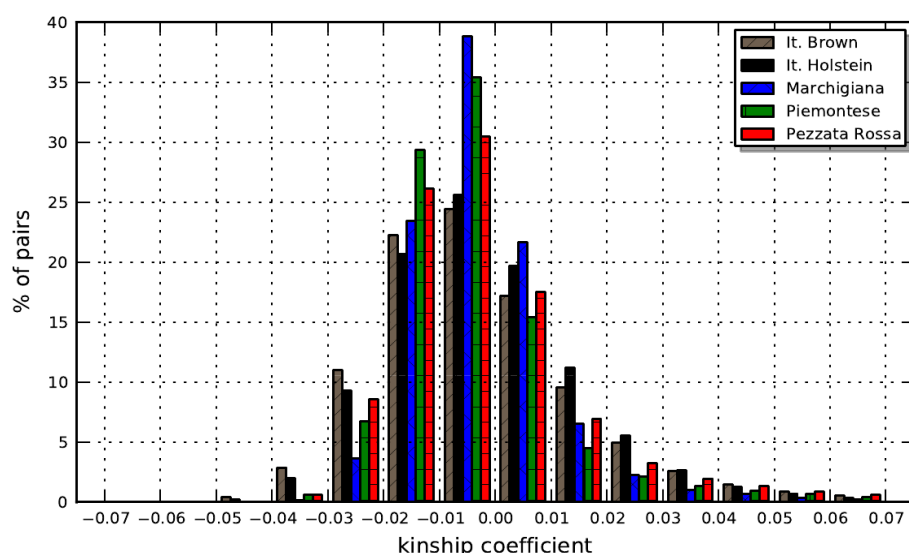


Figure 34: Breed-wise genetic kinship coefficients

The percentage of pairs for a breed is represented in a separate bin on the x-axis. Total percentage of pairs are depicted in grey (Italian Brown), black (Italian Holstein), red (Pezzata Rossa), blue (Marchigiana), green (Piemontese), respectively.

4.3.2.3 Genomic kinship clusters

Figure 34 contains the histograms of kinship coefficients for all possible sire pairs within each breed sample. The percentage of pairs (compared to the population size of each breed) in a given interval is represented as a function of the kinship coefficient. The distribution of values for each sample is centered in the $(-0.01, 0.0)$ interval and negative values are in general more frequent. However the interval $(-0.07, -0.04)$ is unpopulated as compared to its positive counterpart. Piemontese is the breed with greatest negative/positive kinship coefficient ratio, approximately 70% vs 30%; this ratio becomes 65% vs 35% for Marchigiana and Pezzata Rossa and 60% vs 40% for Holstein and Brown. Also, the two dairy breeds show the widest distribution of values while Marchigiana and Piemontese (beef) pairs are concentrated in the central bins and Pezzata Rossa (double purpose) has an intermediate behaviour.

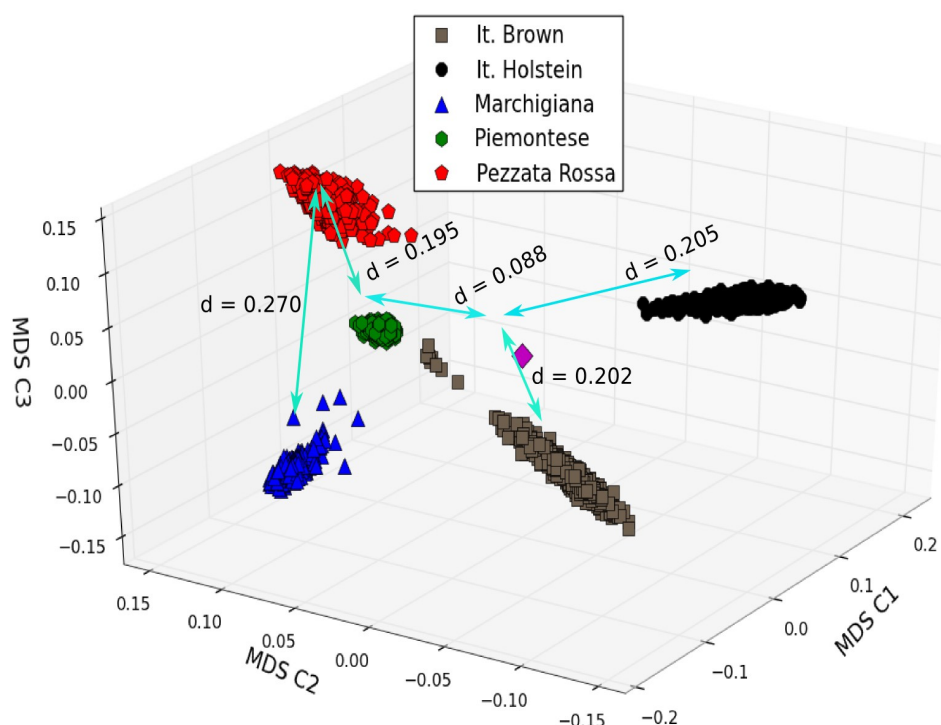
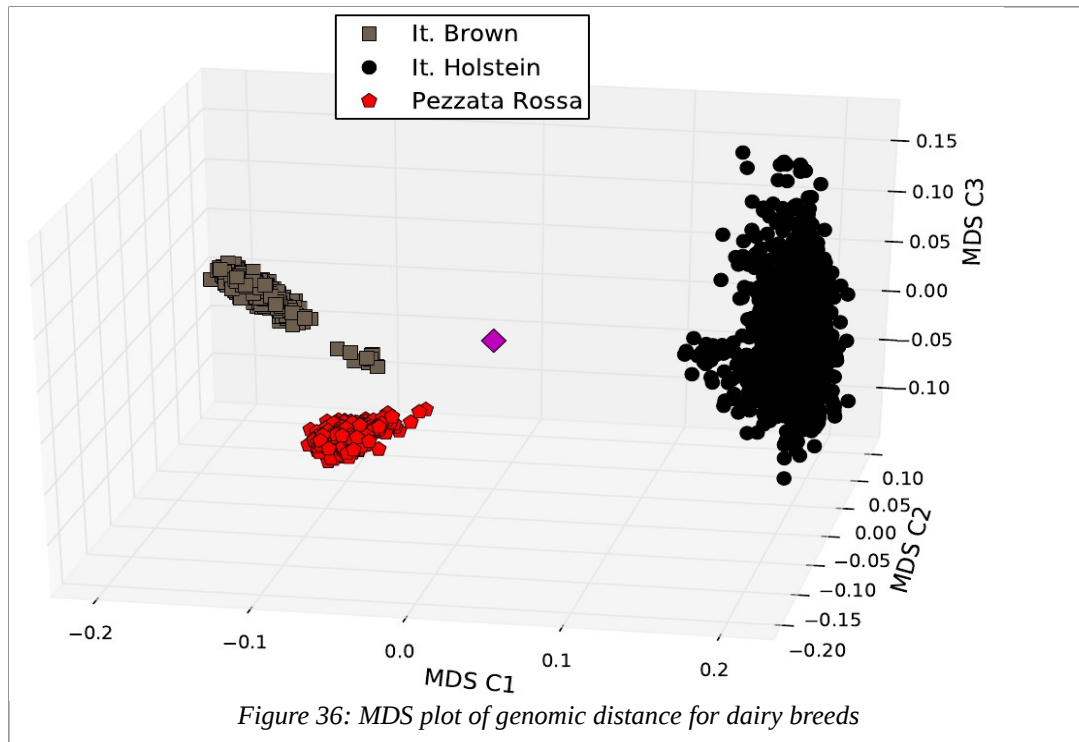
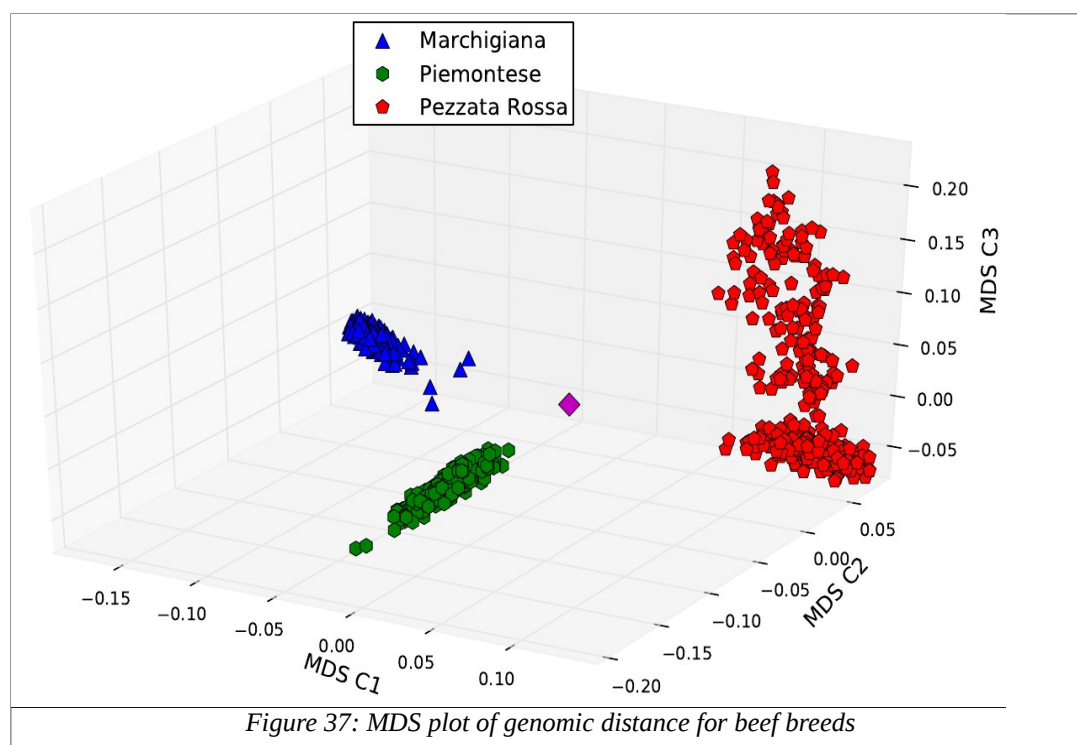


Figure 35: Classical Multidimensional Scaling (MDS) plot of genomic distance for all five breeds. Distance is calculated as 0.5-genomic kinship. The percentage of pairs for a breed is represented in a separate bin on the x-axis. Total percentage of pairs are depicted in grey (Italian Brown), black (Italian Holstein), red (Pezzata Rossa), blue (Marchigiana), green (Piedmontese), respectively.

Figure 35 shows the first three components of the Multidimensional Scaling decomposition of the genetic distance matrix (calculated as 0.5-kinship). Relative distance between selected clusters are shown in the graph and range from 0.195 (Piemontese-Pezzata Rossa) to 0.291 (Marchigiana - Holstein). The five breeds form compact clusters separated from each other with the partial exception of a small group of outlying Brown bulls located near the “centre of mass” of clusters (represented with a diamond). Component one (C1 hereafter) separates Brown and Holstein with Piemontese, Marchigiana and Pezzata Rossa located between these two breeds; Brown and Holstein are both located at comparable distance from the centre on the negative half of component two (C2 hereafter) and separated from the other breeds; both clusters form approximate ellipsoids with the major axis along different diagonals in the plane of the first two components while component three (C3 hereafter) contribution is very small.



The opposite is true for the relative position of beef breeds and Pezzata Rossa for which the greatest differentiation is due to C3. Piemontese forms the most compact cluster and has the shortest distance from the center. Also, it must be observed that, although Pezzata Rossa is a double purpose breed, its cluster is actually more distant from dairy breeds as compared to Marchigiana and Piemontese and that Piemontese is more distant from Marchigiana than from Pezzata Rossa. Figure 36 and 37 show analogous calculations restricted to either dairy or beef breeds. In both cases two breeds (Brown and Pezzata Rossa in Figure 36, Marchigiana and Piemontese in Figure 37) are located in the C1-C2 plane while the remaining population (Holstein or Pezzata Rossa) is located in a different quadrant in the C1-C2 plane and is highly dispersed along C3.



4.3.2.4 Bayesian inference

To estimate the number of genetic clusters within the 2798 cattle samples and 29848 SNPs, a parametric genetic mixture analysis implemented in the STRUCTURE 2.3.2.1 software was performed. Between 2 and 5 clusters (K values) were tested using the admixture model, considering allele frequencies correlated. Consistent results across runs were obtained and a clear clustering of breeds was observed for any K tested. Figure 38 shows assignment probabilities depicted with DISTRUCT for K running from 2 to 5. With K=2 Brown and Holstein clusters are assigned to different populations while sires of the other three breeds belong to each cluster with probabilities near 50%. The separation becomes sharper when three clusters are imposed in the calculation: in this case Marchigiana, Piemontese and Pezzata Rossa sires are assigned to a third cluster different from Brown and Holstein (although some noise is present for Piemontese). With K=4 Marchigiana also forms a distinct cluster while Piedmontese is assigned with comparable probability to either Marchigiana or Pezzata Rossa and more unlikely to Brown or Holstein (the noise present for K=3 is basically conserved). Finally, with K=5 Piedmontese is assigned a distinct cluster.

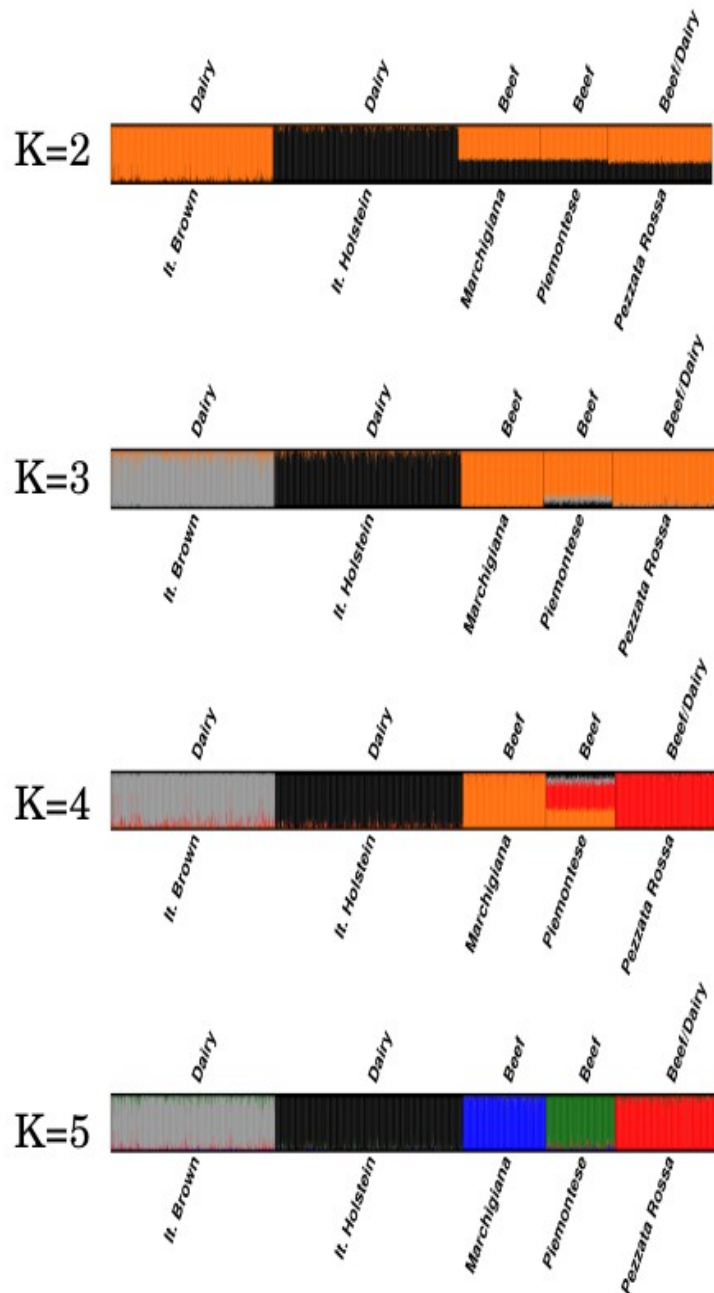


Figure 38: Summary plot of Q estimates for K=2,3,4,5

Estimated membership coefficients for each individual, in each cluster) for K=2,3,4,5 obtained with a 5000 burn-in, and 35000 MCMC, under the admixture model, for the breeds analysed. Individuals are represented in breed groups separated by vertical lines. Breed is indicated above while the breeds' attitude is indicated under each group. Each individual is represented by a single vertical line broken into K coloured segments, with lengths proportional to each of the K inferred clusters. Each colour represents the proportion of membership (M) of each individual (represented by a vertical line) to the K clusters.

4.3.2.5 Fixation Index (F_{ST})

Figure 39 shows F_{ST} values calculated for all autosomal SNPs and averaged over a 8-wide SNP window within each chromosome. Calculation results on all populations, only on dairy breeds and only on beef breeds are displayed in panels A, B and C, respectively. The cutoff values separating the highest and lowest 1% of points are also indicated. In panel A, F_{ST} ranges from 0.04 to 0.3 and average F_{ST} is 0.077 ± 0.058 , with the exception of a single $F_{ST}=0$ signal on BTA19 (Bos taurus autosome 19). SNP clusters with F_{ST} values above 0.0175 (the 99% threshold) are observable on chromosomes 3 to 6, 13, 14, 16, 18, 26 and 29. In most cases these peaks are formed by only one or two groups of 8 markers, therefore spanning on about 4 Mb. More consistent signals are present on BTA13 and on BTA6, where peaks are formed by 11 and 4 groups of markers above the 99% quantile value, respectively. A similar pattern is shown in Panels B and C, where F_{ST} values are calculated for dairy and beef breeds separately; the strongest signals are still observed on BTA6. Interestingly, F_{ST} obtained between dairy breeds are higher than those observed for beef breeds. The peak identified in dairy breeds (figure 4B) spans from 72.35 Mb to 72.47 Mb on BTA6. Three of the four SNPs identified in this region are located in the PDGFRA gene. In beef breeds the peak on BTA6 consists of 6 SNPs spanning from 37.32 Mb to 38.76 Mb (fig. 4C). This interval contains 15 genes including LAP3 (leucine aminopeptidase 3), a gene involved in calving ease. Figure 40A shows F_{ST} values for markers on BTA6 only, without any averaging. Several markers that have F_{ST} above the 99% quantile can be observed at 18 millions of base pairs (Mbp), between 36 and 39 Mbp and at 95 Mbp, but the strongest signal (considering either the number of SNPs with high F_{ST} or the maximum value) is observable at ~38 Mbp. The F_{ST} calculated for dairy breeds (fig. 4B) shows a peak formed by four SNPs spanning from 47,23 Mb to 48,30 Mb on BTA13. In this region we identified 14 genes among which CDS2 could affect milk fat composition. Four SNPs with high F_{ST} were also identified on chromosome 13 at positions 67197635, 67464116, 67490718, 67766784 when considering beef breeds (fig. 4C). Of the 8 genes in this region, SRC and CTNNBL1 may be a possible candidate genes. Figures 5B and 5C show inbreeding coefficients (F_{IS}) calculated within each breed. No relevant peaks of F_{IS} values are observable at the coordinates of the F_{ST} peaks in Figure 5A.

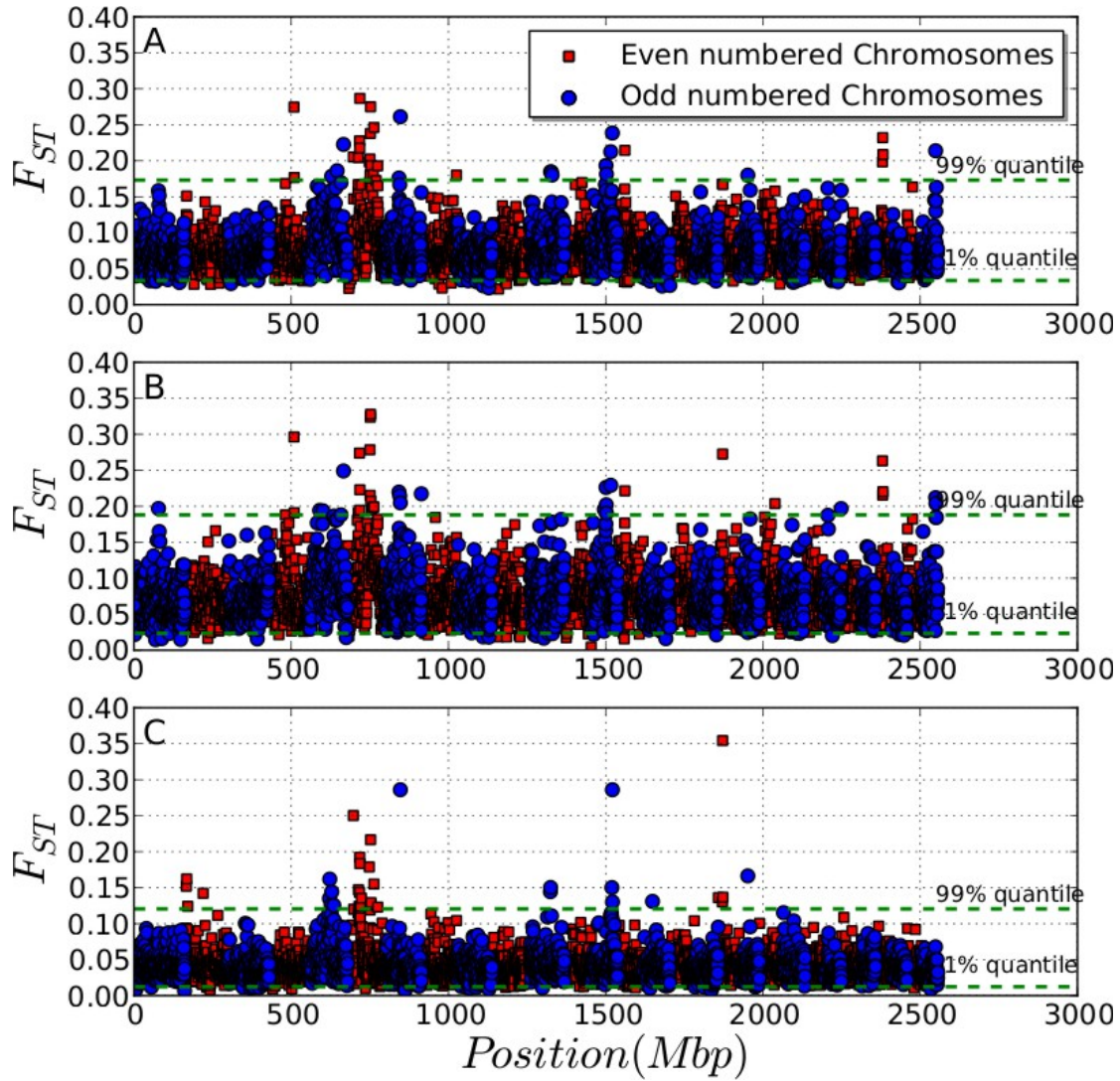


Figure 39: Genome wide Manhattan plots of Fixation index (F_{ST})

Fixation index (F_{ST}) has been calculated for all polymorphisms averaged over a window of 8 SNPs within each chromosome and calculated for all five breeds (Panel A), dairy breeds (Panel B) and beef breeds (Panel C). F_{ST} for odd and even chromosomes are represented with blue circles and red squares, respectively. The values corresponding to the lowest 1% and 99% quantiles are represented with green dashed lines.

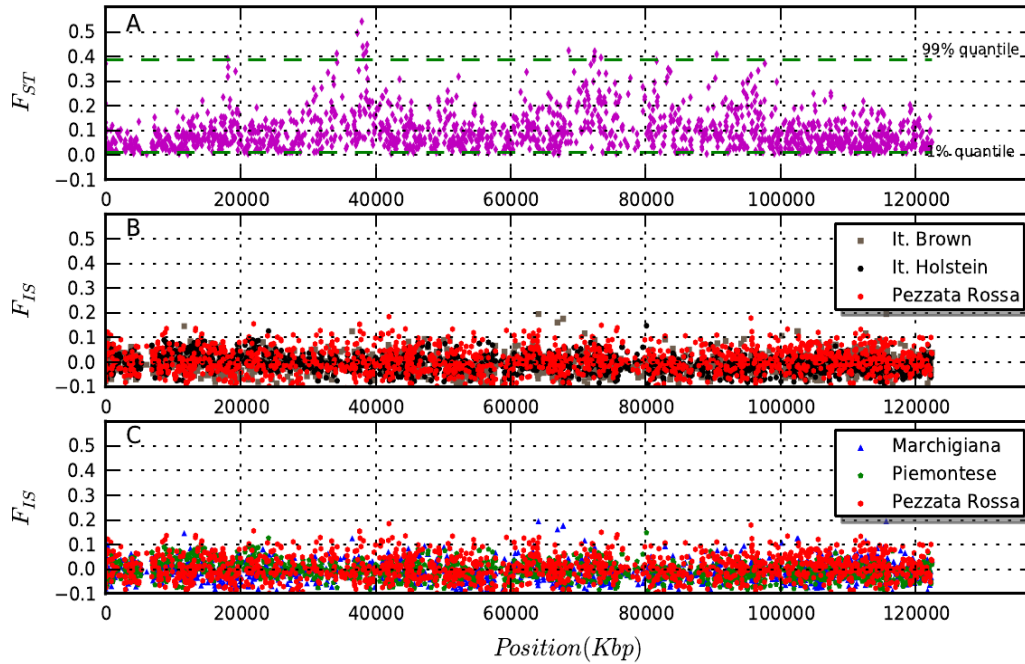


Figure 40: F_{ST} calculated for all markers on BTA6

F_{ST} calculated for all markers on BTA6 averaged over a 3 SNP window. The values corresponding to the lowest 1% and 99% quantiles are represented with green dashed lines. Panel B: Plot of inbreeding coefficients F_{IS} calculated for all markers on BTA6 within each dairy breed. Values for Brown, Holstein and Pezzata Rossa are depicted with a gray line, a black dashed line and a red dot-dashed line. Panel C: Plot of inbreeding coefficients (F_{IS}) calculated for all markers on BTA6 within each dairy breed. Values for Marchigiana, Piemontese and Pezzata Rossa are depicted with a blue line, a green dashed line and a red dot-dashed line

4.3.3 Discussion

In recent studies high-density panels of SNP data have been used to examine population structure among cattle breeds and to cluster them into distinct groups (zebu, taurine, or hybrid breeds) ([104], [105], [107], [115]). In this study we investigated the potential of SNP markers to assess the genetic structure of 5 Italian cattle breeds using the BovineSNP50 BeadChips in combination with Bayesian and multidimensional scaling approaches. The clustering analysis revealed the complete genetic isolation and the absence of admixture in the five breeds considered. These results reflect well the characteristics of the considered breeds and the impact of the selection on their genetic structure. In particular the multidimensional scaling separated each breed in five well defined clusters. Piemontese forms the most compact cluster indicating that the selection activity and the extensive use of artificial insemination was more intensive in this breed. It is worth to observe that the double purpose Pezzata Rossa is more distant from dairy breeds than beef breeds, suggesting in this breed a different management

of selection and lack of admixture with other dairy breeds. The genetic isolation and lack of admixtures among the five breeds are confirmed by Bayesian analysis in which the breeds are clustered on the basis of their attitudes. This means that the anthropic selection in cattle has led to a genome divergence between the beef and dairy breeds. Moreover, the 29848 SNPs utilized allowed to assign individuals to the breed of origin with a probability of 0.94% which is higher than those reported in previous studies where microsatellite markers were used ([116], [117]) but comparable to values obtained by Ciampolini et al. and Negrini et al. [100], [118].

After analysing the genetic structure of the studied populations, we presented an application of selection signatures in the same samples. Mapping the F_{ST} values against genome location we identified genes with signatures of positive selection involved in the biological processes such as reproduction, metabolism of lipids, calving ease. A strong selection signal was observed on BTA6 when considering F_{ST} across all cattle breeds. Interestingly, the F_{ST} calculated only in dairy breeds revealed the evidence for selection in the region located at 72.45 Mb on chromosome 6, within the PDGFRA gene. PDGFRA is associated with β -estradiol which is implicated in the reproduction process and in the regulation of calcium level, a molecule involved in milk metabolism. These results are consistent with those obtained by Flori et al. [119] who found a positive selection signature in the same region in Holstein cattle breed. Instead, Stella et al. [95] proposed a KIT gene as responsible of selection signature in this region among dairy breeds.

The peak observable at ~38 Mbp on BTA6 in beef breeds was localized near LAP3 which can be considered as possible candidate gene. LAP3 encodes for a leucine aminopeptidase, which is responsible of the oxytocin hydrolysis ([120], [121]). Recent studies demonstrated the role of LAP3 in calving ease in Norwegian Red cattle [122]. A cluster of signals reflecting strong evidence of selection was also observed in BTA13. When we analysed separately the F_{ST} for beef breeds (Figure 39C) a strong peak at position 67 Mbp can be observed. Two interesting genes are located in this region: SRC and CTNNBL1. SRC is involved in the regulation of actin cytoskeleton and in the focal adhesion pathway. Some studies reported the role of focal adhesion pathway for muscle formation in cattle [123] and muscle strength and integrity in racing horses [124]. The other gene, CTNNBL1 catenin, is associated with body weight and height [125]. In human it has been showed to be involved in the Wnt/beta-catenin-sig-

naling pathway and associated with obesity. In dairy breeds we observed a peak spanning 47-48 Mbp on BTA13 near the CDS2 gene which could influence milk fat composition; the gene in fact is involved in the phospholipid biosynthetic process.

4.4 Identification of a short region in chromosome 6 affecting calving traits in Piedmontese cattle

4.4.1 Introduction

Reproductive health in livestock is an economically important trait and calving is a key element to improve productivity and animal welfare. Difficulties in calving have an evident negative impact on the sustainability and profitability of a herd, by adding considerable costs for veterinary treatments and forcing slower re-mating. Calving in cattle is affected by calf traits and morphology and by dam characteristics ([126], [127]) and can be described by two traits: maternal calving ease and direct calving ease. Both refer to calving without human intervention, the first attributed to the mother and the second to the calf. Calving ease is affected by many environmental and genetic factors. In addition to birth weight, other important factors as calf shape, the size [128–133] of cow pelvic area or the gestation length are involved [134]. One of the major causes of dystocia is a disproportion between size of the calf at birth and the cow birth canal. Thus, the cow physiological predisposition to birth and the calf morphology are the key genetic factors which can be improved by breeding plans [135]. The Estimated Breeding Values (EBVs) for calving ease used in breeding plans are measured from three sets of records: calving ease score, birth weight and gestation length. Several studies have identified QTLs influencing calving traits and a total of 53 QTLs for calving ease are reported in public databases (http://www.animalgenome.org/cgi-bin/QTLdb/BT/qtraitology?trait_ID=1073). The estimation of EBV and the identification of genes significantly associated with calving performance is a challenging task since the traits show low heritability and the recording is made on scales which are subject to breeder's interpretation. The recent availability of cattle genome-wide SNP panels could potentially overcome these drawbacks [136]. The aim of this study is to identify cattle genome regions possibly affecting direct and maternal calving performance in the Piedmontese breed, using a genome-wide association study. This Italian breed carries muscular hyperplasia [137] and has been selected for beef production. In this work, a medium density SNP marker panel (54,001 SNPs) was used

to investigate the genome of 323 Piedmontese bulls for identifying chromosomal regions associated with calving traits. This genome-wide scan let us to discover a 1.8 Mb region of chromosome 6, associated with direct calving ease.

4.4.2 Results

4.4.2.1 Preprocessing

Sires and markers with a call rate under 95% were discarded; as well as SNPs having a minor allele frequency (MAF) under 2.5%. Sires were checked for abnormally high autosomal heterozygosity (FDR 1%) [86]. Hardy Weinberg Equilibrium was checked on the filtered data set [109]; markers were removed from the data set if the P-value was less than 0.01. Kinship among sires was estimated directly from genomic data as proposed by Astle and Balding [62]. Three of the 323 sires were excluded having a genotype call rate lower than 95% and two for having a too high autosomal heterozygosity. As for markers, 506 out of the 54,001 SNPs of the chip were excluded because the ratio of missing genotypes was higher than 5%; 10870 were excluded as showing minor allele frequency below 2%; 1372 because they were not in HWE ($P < 0.01$). The final dataset was thus formed by 319 sires and 41333 SNPs. Figure 1 shows the distribution of EBV values for direct calving ease partitioned in five equally spaced classes; an acceptable approximation of a normal distribution can be observed for direct calving ease, with the three central intervals accounting for 75% of sample size; a tiny extreme negative tail which is not mirrored by positive values can be observed.

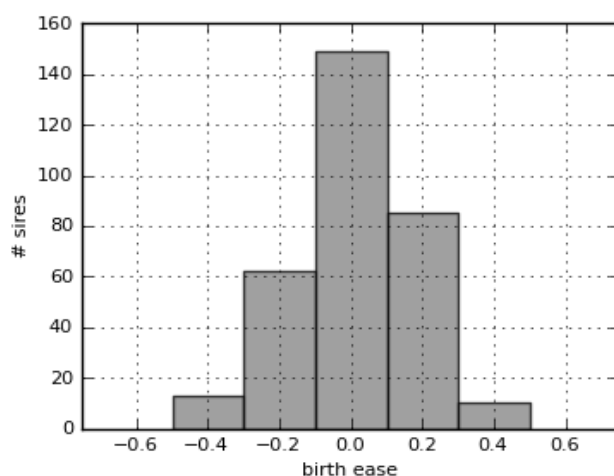


Figure 41: Histogram of EBV values for direct calving ease

4.4.2.2 Genome-wide scan

Association of SNPs to phenotypes was performed using the Family based Test for Association to reduce computing time, proposed by Chen and Abecasis [66]. Kinship among sires was treated as a random effect, while SNP genotypes and age were considered as fixed effects. The model used was:

$$y = 1\mu + Zu + u_g g + e \quad (19)$$

where μ is the mean and 1 is a vector of ones, Zu accounts for polygenic effects, u_g is the effect of the marker under investigation, g denotes a genotype, and e is a vector of residuals. Polygenic effects were estimated by calculating genomic kinship between sires using the complete set of filtered SNPs [56], [57]. Genomic control was applied and the inflation factor λ was calculated and used to correct significance levels [138]. False Discovery Rate control was performed using the Benjamini-Hochberg method using a FDR of 0.01. Linkage Disequilibrium (LD) in selected regions was estimated calculating the r^2 coefficient between SNPs using the approach proposed by Hao [139]. Results of the genome-wide scan for direct calving ease traits are summarized in Figure 2A: SNPs are colored chromosome-wise and Log significance levels for the association between SNP and traits are shown as a function of marker positions concatenated chromosome length. Among the 41333 polymorphisms, 2030 (roughly 0.5%) were associated to direct calving ease. A very sharp peak containing several SNPs with very high $-\log_{10}(P^*)$ values can be observed on the first half of BTA6, between 37 and 39 millions of base pairs (i.e. between 715 and 720 Mbp relatively to total genome length). The inflation factor λ obtained for the test was 1.03 showing that, given the number of markers, residual inflation was acceptable. Correction for λ yielded a total of 1876 significant SNPs; of these, 1833 had a λ -corrected P-value in the interval $0.05 > P \geq 0.001$ and 43 had a λ -corrected p-value of less than 0.001. Figure 42 shows SNP association to direct calving ease after the FDR control: thirteen SNPs, delimiting the previously described association peak on BTA6 (see Figure 42), resulted to be significantly associated to direct calving ease. Significance levels (unadjusted, P^* , and adjusted with the Benjamini-Hochberg method, Q^*) of these SNPs, together with the SNP name and position on BTA6 according to Illumina 54K Beadchip specifications, and the gene in which they are located (if any) are shown in Table 7.

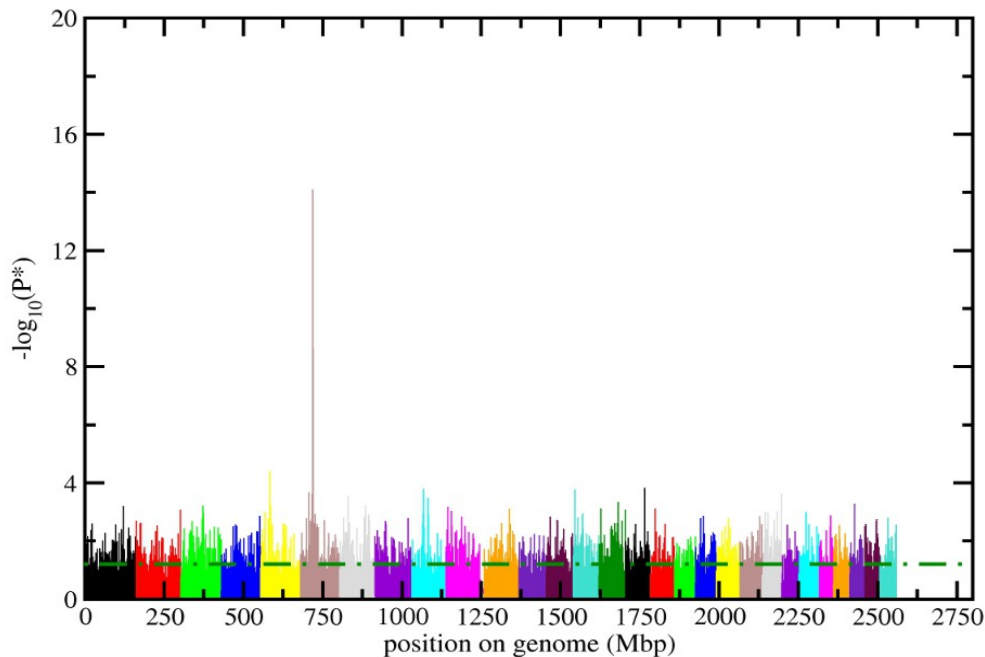


Figure 42: Manhattan plot of Log significance levels of SNPs associated with direct calving ease

Results for the entire filtered dataset before the correction for inflation are shown. Values above the dot-dashed green line denote a $P^* < 0.05$.

Several markers have a P-value $< 10^{-7}$ after the correction for multiple testing and a clear association of this chromosomal region with the studied trait can be observed. Allelic effects of the minor allele as compared to major (last column of Table 7) are small and positive except for two markers (Hapmap26308-BTC-057761 and ARS-BFGL-NGS-45457). Figure 43 shows the region of BTA6 spanned by these 13 SNPs, the genes located in the region and the position of the SNPs listed in Table 7; one marker (#2) is within the sequence of the LAP3 gene while four (SNPs #5, #6, #7, #8) are located in the LCORL gene.

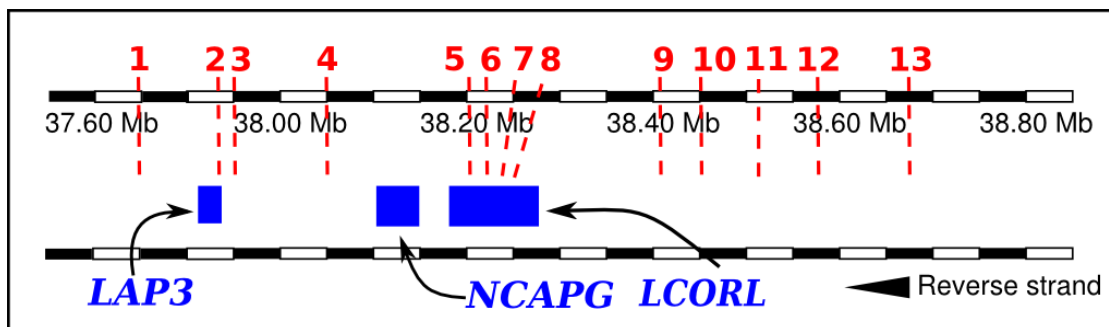


Figure 43: Genes in the BTA6 peak

The approximate position and size of genes located in the region (blue rectangles) is shown together with the position of each SNP (red numbers and red dashed lines).

4. Cattle quantitative genetics

Figure 44 shows r^2 estimates of pairwise Linkage Disequilibrium for all polymorphisms located in this region: a high correlation between genotype frequencies can be observed even for distant SNPs. Average r^2 was 0.426 when all SNP pairs were considered and comparable values when obtained for SNPs within versus outside of genes (0.413 and 0.416 for LAP3/NA and LCORL/NA, respectively) but higher correlation was observed between loci within LCORL ($r^2 = 0.637$) or between LCORL and LAP3 ($r^2 = 0.571$).

SNP #	SNP name	Position (bp)	gene	P*	Q*	A	B	fB	β B
1	Hapmap30134-BTC-034283	37,852,401	NA	2.38E-09	1.94E-05	C	T	0.45	8.43E-02
2	Hapmap26308-BTC-057761	37,963,148	LAP3	8.17E-15	3.38E-10	G	A	0.40	-1.19E-01
3	ARS-BFGL-NGS-112812	38,014,255	NA	3.87E-06	1.60E-02	G	A	0.35	6.85E-02
4	ARS-BFGL-NGS-45457	38,102,328	NA	7.97E-13	1.65E-08	T	A	0.45	-9.97E-02
5	Hapmap27083-BTC-041166	38,212,942	LCORL	2.58E-07	1.52E-03	A	G	0.36	7.50E-02
6	Hapmap23507-BTC-041133	38,233,089	LCORL	3.89E-07	2.01E-03	C	A	0.36	7.39E-02
7	Hapmap31285-BTC-041097	38,256,890	LCORL	4.11E-11	4.24E-07	T	C	0.48	-9.30E-02
8	Hapmap33628-BTC-041023	38,326,148	LCORL	7.12E-12	9.80E-08	A	G	0.46	-9.85E-02
9	Hapmap25414-BTC-034877	38479644	NA	2.45E-05	7.78E-02	G	A	0.27	6.71E-02
10	Hapmap32207-BTC-034871	38,500,210	NA	2.45E-05	7.78E-02	A	G	0.27	6.71E-02
11	Hapmap28546-BTC-072715	38,558,527	NA	4.84E-07	2.22E-03	C	G	0.38	7.38E-02
12	Hapmap27537-BTC-060891	38,638,963	NA	1.04E-05	3.90E-02	C	T	0.39	6.68E-02
13	Hapmap31044-BTC-071337	38,729,867	NA	2.38E-09	1.94E-05	C	T	0.46	8.66E-02

Table 7: SNPs associated to calving ease from GWAS

Single nucleotide polymorphisms associated to direct calving ease and located on BTA6 in the region comprised between 37 and 39 millions of base pairs sorted according to their position on the Illumina panel. SNP name, position, the known region in which they are located (if any), significance levels both unadjusted (P^*) and corrected for multiple testing (Q^*), allelic frequencies (f_B) and effects of the minor allele (β_B).

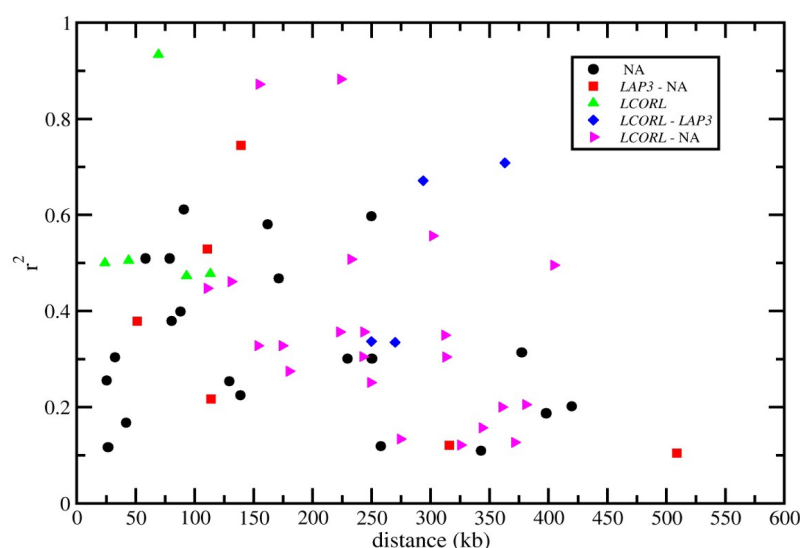


Figure 44: Pairwise LD between SNPs on BTA6 peak

SNP pairs are coloured by the location of one or both markers for which r^2 is calculated. SNPs not in the *LAP3* or *LCORL* genes are marked as “NA”.

4.4.2.3 SNP discovery, genotyping and analysis

Since the genome wide analysis pointed to a ~1.8 Mb region associated to direct calving ease (see Results) we searched for further single nucleotide polymorphisms in genes included in this region to validate and enrich the results of the genome wide scan. We selected eight SNPs in the *LAP3* gene and four in the *NCAPG* gene (Table 1). For *LAP3* gene three markers were described by Zheng [121], while four SNPs were selected by screening *in silico* NCBI dbSNP (<http://www.ncbi.nlm.nih.gov/snp/>, <http://www.ncbi.nlm.nih.gov/projects/SNP/>) and one was present in the Illumina 54K. The four SNPs in *NCAPG* gene were described by Setoguchi [140], [141]. The 12 SNPs were genotyped on 20 bulls selected accordingly to extreme EBV values for direct calving performance (10 in each tail). Primer pairs (Table 8) were designed using the primer3 software (<http://fokker.wi.mit.edu/primer3/>) and synthesized by Sigma-Aldrich. Each fragment was amplified with a 1x Fidelity PCR Master Mix (Stratagene, Agilent) in a volume of 50µl containing 50ng of genomic DNA and 1pMol of each primer. The amplification was carried out on a PTC-100 (MJ Research) following standard conditions. PCR products were purified through ChargeSwitch (Invitrogen) and outsourced to MacroGen Inc. (www.macrogen.com, <http://dna.macrogen.com/eng/>) for sequencing. Sequences were analysed using the BioEdit software [142].

SNPs in LAP3	Forward	Reverse	Ta (°C)	PCR p. s. (bp)	Reference	region**
rs110839532	AGGGAAGGAGA TGGACAGTATG GTC	AGAGCCACACC AGAGAGCACCA	60	810	NCBI dbSNP	3'UTR
rs43702361	"	"	60	810	NCBI dbSNP	3'UTR
rs109241256	"	"	60	810	NCBI dbSNP	3'UTR
rs41255599	"	"	60	810	NCBI dbSNP	3'UTR
rs43702364	GTCATTACCAC AAGAAAA	AATGTGCCAC TTAGGAT	45	278	Zheng et al., 2010	intron 12 (24428-24878)
rs43702363	"	"	45	278	Zheng et al., 2010	intron 12
rs43702362	"	"	45	278	Zheng et al., 2010	intron 12
rs109315366	TCTGGATTCTT GGCCATTTTC	GCCAACTCATA GGCTGGAAG	57	284	Illumina panel	Intron 1
SNPs in NCAPG	Forward	Reverse	Ta (°C)	PCR p. s. (bp)	Reference	
rs109570900	TTTCAGAATGT GAATTTTGGCT TA	AGCCAAAAGCA CTGAAAACAC	56	370	Setoguchi et al. 2009	exon 9
NA*	TTCTGTATGTG CCCGACTGA	TCTGGCAGCTA AATTAAGCAAA	56	400	Setoguchi et al. 2009	intron
rs110251642	GAGCTTACATG GGGAGGGTTA	CTTCAAGAAAT GAGCACAAA	57	384	Setoguchi et al. 2009	exon 17
NA*	TTTAAAGCTCAA TGGAAATCAGGA	TGGAATCGCAC ACCAGAAAT	57	357	Setoguchi et al. 2009	intron

Table 8: Additional SNPs sequenced in the LAP3 and NCAPG genes

The rs number, forward and reverse sequences, temperature of annealing, PCR product size and gene region are shown.

*rs was not available; see the listed reference for complete information

**according to Ensembl Bos taurus version 63.4 (Btau_4.0)

Association between these SNPs and trait values for direct calving ease was assessed using a simple linear regression model. Based on the results of genome wide scan we decided to genotype further SNPs in genes located in the region where we found a strong association with direct calving ease for which the Illumina panel was less rich of markers. We selected eight SNPs in *LAP3* (only one marker from the Illumina panel was located in this gene) and four in *NCAPG* (not represented in the Illumina panel). SNPs #14 to #19 (see Table 9) were in complete Linkage Disequilibrium ($r^2 = 1$) and shared the same high association with direct calving ease; a high positive effect is associated with the major allele ($P^*=6.947E-05$, adjusted $R^2=0.572$, $\beta=0.395$). SNPs #20 was the only marker in *LAP3* not significantly associated with direct calving ease. SNP #21 showed a strong but not complete LD with markers #14 to #19 ($r^2=0.795$); results were consequently slightly different: $P^*=2.910E-04$, adjusted $R^2=0.500$ and $\beta=0.378$. Regarding SNPs analysed in the *NCAPG* gene, two were weakly associated to direct calving ease showing a positive effect associated with the major allele: rs109570900 ($P^*=0.010$, $R^2=0.276$, $\beta=0.279$) and rs110251642 ($P^*=0.020$, adjusted $R^2=0.223$, $\beta=0.268$). Average Linkage Disequilibrium estimation for all 13 SNPs was $r^2=0.551$,

comparable to that observed between *LAP3* and *LCORL* although estimated on a much smaller sample.

		Extreme negative Sires										Extreme positive Sires									
#SNP	rs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		Genotypes																			
14	rs43702364	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
15	rs43702363	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
16	rs43702362	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
17	rs110839532	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
18	rs109241256	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
19	rs41255599	1	1	2	2	1	1	1	2	1	2	1	0	0	0	2	1	1	0	2	2
20	rs41255598	2	2	0	1	1	1	2	1	1	0	2	2	2	2	0	2	1	2	2	1
21	rs109315366	1	2	2	2	1	1	1	2	1	2	1	0	0	1	2	2	1	0	2	2

Table 9: SNPs in the *LAP3* gene

SNP number, SNP name and genotypes for markers selected in the LAP3 gene. Genotypes are coded as 0,1,2; 0 stands for minor allele homozygote, 1 for the heterozygote and 2 for the major allele homozygote. Columns in italics indicate sires with extreme high direct calving ease EBV value

4.4.3 Discussion

Today, selection objectives in agriculture pay much attention to robustness, longevity, welfare and, more generally, sustainability. Calving ease traits are aimed at these purposes, have a large impact on animal health and farm profitability, and they also make feasible to breeders the application of livestock systems [130]. However, antagonistic relationships between calving performance and beef production traits in Piedmontese [143] lowers the usefulness of classical breeding strategies and highlights the need for specific selection strategies. The possibility to select breeding individuals based on genotypes at genes affecting calving performances may help to overcome this problem. Piedmontese is a breed carrying the G1056 mutation in the myostatin gene ([137], [144]). As for other mutations in myostatin, calving ease is affected although less severely than in other affected breeds (e.g. Belgian Blue). Estimates of heritability of calving related traits for the Italian Piedmontese population yielded values for the maternal and direct effects between 0.19 and 0.08 depending of the number of parities and the type of effect considered [145]. suggestive QTLs on various chromosomes, affecting calving ease and foetus growth-related traits in cattle are described: Pausch and collaborators identified two major loci on BTA 14 and 21 [136]. Several QTLs for

calving ease are reported in public databases; In particular, different studies point to the existence of QTLs for parturition and foetal growth on chromosome 6 in various cattle populations ([146–148]). We performed a genome wide scan of Piedmontese cattle breed in Italy, using a medium density chip featuring about 54K SNPs. Results showed a cluster of 13 SNPs in a 1.8 Mb wide region of BTA6 all significantly associated ($P^* < 0.001$) with direct calving ease. Three interesting genes are located in this region: LAP3, encoding for a leucine aminopeptidase, which is responsible of the oxytocin hydrolysis ([120], [121]); NCAPG, encoding for a non-SMC condensin I complex, which has been associated, in cattle, with foetal growth and carcass size ([140], [141]); LCORL, encoding for a ligand dependent nuclear receptor corepressor-like, which affects human height and also controls stature in cattle in high correlation with NCAPG [63].

Estimation of LD showed that markers located in LAP3 (one marker) and LCORL (four markers) were correlated but unfortunately no SNPs from the Illumina panel were found within NCAPG. Thus, we enriched the SNP set for LAP3 and NCAPG, genotyping a subset of sires with extreme EBVs; eight and four SNPs were selected in LAP3 and NCAPG, respectively. Of these, seven in LAP3 and two in NCAPG were significantly associated with the trait; moreover the markers in LAP3 were in complete LD, pointing to the existence of a possible selection signature, to be confirmed by analysing a larger number of animals. Thus, the results of additional genotyping fully confirmed the findings of the genome-wide scan. LAP3 gene, being involved in oxytocin metabolism and hence in maternal calving effects, may not influence direct calving ease. In this case, its alleles may be in Linkage Disequilibrium with the alleles of the other two relevant genes in the region.

It is well known that the modulation of genes implied in foetal growth is often under epigenetic parental control (genetic imprinting). In general, the paternal allele of imprinted genes up-regulates the foetal growth whereas the maternal one acts in the opposite way (“Parental tug of war” [149], or “Sex conflict”), in order to preserve its reproductive efficiency. According to our results, we can figure out that NCAPG, and probably LCORL, genes are involved in the regulation of foetal growth and direct calving ease and that they could be under imprinted parental control. Following our hypothesis the paternal genotype at these two loci should limit or increase the positive influence of these genes on foetal growth. Recently it has been shown that a poly-

morphism in the NCAPG gene (rs109570900, 1326T>G) is significantly associated to cell division [150], to adult human height ([151], [152]) and to an increase in bovine carcass weight at puberty ([140], [141]). Our analysis reveals a weak association of both rs109570900 and rs110251642 to direct calving ease and a positive effect linked to the homozygote genotypes for the major allele (T/T in rs109570900 and C/C in rs110251642). These results confirm the existence of an antagonistic relationships between calving performance and beef production traits [143].

5 Summary of results

5.1 Water buffalo genome assembly

Water buffalo is a fundamental asset to the lives of small farmers and to the economy of many countries worldwide in particular in developing countries in South East Asia. As the results on cattle presented in this work shows, modern genomics techniques may lead to a variety of biological results and potential animal breeding applications. Since any almost any genomic investigation is based on the availability of a reference genome the importance of the water buffalo genome assembly that is still being completed is clear. Preliminary results of the guided assembly confirms the high similarity between water buffalo and cattle which may be of fundamental importance to further advance water buffalo genomics exploiting vast amount of cattle genetic/genomic resources.

5.2 Candidate genes in Italian Brown

Candidate genes were selected on the basis of known relationships between physiological or biochemical processes and on evidence of their significant association with milk production traits in various species. Seven SNPs were found to be associated to one or two traits. Of these, CSN1S1_26181 already showed intensive selection of the A allele while the remaining SNPs had comparable allelic frequencies even when a less common genotype was present (e.g. TT for LGB_5261 and THRSP_rs42714483); these polymorphisms can thus be considered as selection targets, particularly when alleles associated to high DYD values have a suitable frequency in the population. On these premises, OXT and can be considered a good GAS (Gene assisted selection) candidates for milk yield, and LGB for both milk and protein yield. LGB genotypes positively associated to milk and protein yield are already quite common in the analysed population, but their frequency could be further increased.

Summarising, we proved association between six of the analysed genes and the selected traits in Brown breed. Since the actual usefulness of gene/traits associations in artificial breeding must be tested in any population studied, our work provides useful information that could be used for selection purposes in Italian Brown breed. ANARB already includes k casein alleles in GAS programs: the identified alleles present at

suitable frequency could be included in ongoing selection programs affecting polygenic traits. The findings of this study indicate that both SNPs within LGB are a good candidates deserving further detailed studies on their functional mechanisms.

5.3 Assessment of population structure in five Italian cattle breeds

Recently many innovative tools, such as medium or high density SNP chips, have been developed for various domesticated species. Due to their abundance and to the availability of high throughput analysis technologies, SNP markers are replacing microsatellite loci and have begun the most suitable markers for population genetic analyses in cattle. Using a high density SNP panel we could assess the genetic structure of 5 Italian cattle breeds. Both MDS and Bayesian clustering indicate the effects of anthropic selection, particularly revealing a strong genetic divergence between beef and dairy breeds. Moreover we could identify selection signatures peculiar of each breed type, which suggested selection for genes involved in milk or meat traits. We will further study these regions and the genes harboured in these regions, trying to associate genetic variations with breed characteristics.

5.4 Identification of a short region in chromosome 6 affecting calving traits in Piedmontese cattle

Our genome-wide association study shows the power of EBVs in resolving the genetic complexity of heritable traits. The analysis of 323 bulls by a medium density SNP marker panel revealed that several SNPs were significantly associated with direct calving ease. A very strong signal was detected in a region comprised between 37.8 and 38.7 Mb on BTA6. Five of the markers in this region were within genes: one in the leucine aminopeptidase 3 (LAP3) gene, and four in the ligand dependent nuclear receptor corepressor-like (LCORL) gene. Linkage Disequilibrium analysis showed high correlation between SNPs located in LAP3 and LCORL as compared to all significant SNPs. Then, we genotyped additional SNPs within in two subsets of bulls showing extreme EBVs; eight SNPs were genotyped in LAP3 and four in non-SMC condensin I complex, subunit G (NCAPG) for which no polymorphisms were present in the SNP marker panel; the results of this additional association analysis fully confirmed the results of the genome-wide scan. Minor allele frequencies for significant SNPs ranged from 0.25 to 0.45 meaning that all SNPs can be considered as selection targets to improve the frequency of the favorable allele.

5.5 Conclusions

This work gives a glimpse of the impact that current (and future) sequencing and bioinformatics techniques have on biology. Starting from shotgun sequencing a draft genome may be obtained; comparing variations to this reference allows to unravel genetics diversity between populations of the same species or subspecies, and to link genetic information to almost every kind of trait information available. The ever decreasing cost of sequencing allows for continuous enlargement of samples and for improvement of references and SNP panels. Also, with decreasing costs, several sequencing projects for farm or wild (and maybe endangered) species may become feasible. Genomics applications to animal (or plant) breeding may have drastic consequences on several fundamental issues for world economics and politics given the impact that easy and convenient genetic improvement of farm animals may have on the necessity to feed an (increasing) world population of seven billion people in a sustainable way: genetically improved animals may be able to feed in poorer and/or different environments, may be more resistant to diseases or require less care by man and finally they may be able to yield better products in greater quantities.

6 References

- [1] J. D. Watson, *Molecular biology of the gene*. San Francisco; Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ; Cold Spring Harbor Laboratory Press, 2008.
- [2] G. Sutton and I. M. Dew, "Shotgun Fragment Assembly," in *System Biology: Genomics*, Oxford University Press, 2007.
- [3] M. Pop, S. L. Salzberg, and M. Shumway, "Genome sequence assembly: algorithms and issues," *Computer*, vol. 35, no. 7, pp. 47-54, Jul. 2002.
- [4] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen, "Nucleotide sequence of bacteriophage λ DNA," *Journal of Molecular Biology*, vol. 162, no. 4, pp. 729-773, Dec. 1982.
- [5] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick, "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd," *Science*, vol. 269, no. 5223, pp. 496-512, Jul. 1995.
- [6] S. Zhao and M. Stodolsky, *Bacterial artificial chromosomes*. Humana Press, 2004.
- [7] S. M. D. Goldberg, J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y.-H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter, "A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes," *Proceedings of the National Academy of Sciences*, vol. 103, no. 30, pp. 11240 -11245, Jul. 2006.
- [8] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A Whole-Genome Assembly of *Drosophila*," *Science*, vol. 287, no. 5461, pp. 2196 -2204, Mar. 2000.
- [9] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson,

- L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304-1351, Feb. 2001.
- [10] E. R. Mardis, "Next-generation DNA sequencing methods," *Annu Rev Genomics Hum Genet*, vol. 9, pp. 387-402, 2008.
- [11] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nat. Methods*, vol. 5, no. 1, pp. 16-18, Jan. 2008.
- [12] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg, "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, no. 7189, pp. 872-876, Apr. 2008.
- [13] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson, "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Res.*, vol. 18, no. 7, pp. 1051-1063, Jul. 2008.
- [14] G. J. Porreca, "Genome sequencing on nanoballs," *Nat. Biotechnol.*, vol. 28, no. 1, pp. 43-44, Jan. 2010.
- [15] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcharding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, and C. A. Reid, "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays," *Science*, vol. 327, no. 5961, pp. 78-81, Jan. 2010.

- [16] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth, and J. Bustillo, "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, no. 7356, pp. 348-352, Jul. 2011.
- [17] M. Pop, "DNA sequence assembly algorithms," in *2006 Yearbook of Science and Technology*, McGraw-Hill, 2005.
- [18] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 354-366, May 2009.
- [19] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315-327, Jun. 2010.
- [20] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander, "ARACHNE: A Whole-Genome Shotgun Assembler," *Genome Res*, vol. 12, no. 1, pp. 177-189, Jan. 2002.
- [21] X. Huang and S.-P. Yang, "Generating a genome assembly with PCAP," *Curr Protoc Bioinformatics*, vol. Chapter 11, p. Unit11.3, Oct. 2005.
- [22] S. Batzoglou, "Algorithmic Challenges in Mammalian Genome Sequence Assembly," in *Encyclopedia of genomics, proteomics and bioinformatics*, John Wiley and Sons, 2005.
- [23] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment," *J. Comput. Biol.*, vol. 1, no. 4, pp. 337-348, 1994.
- [24] J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, and G. Sutton, "Aggressive assembly of pyrosequencing reads with mates," *Bioinformatics*, vol. 24, no. 24, pp. 2818-2824, Dec. 2008.
- [25] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 9748-9753, 2001.
- [26] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821-829, Feb. 2008.
- [27] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABYSS: A parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117-1123, Feb. 2009.
- [28] R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. A. Ryder, F. C.-C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T.-Y. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T.-W. Lam, S.-M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K.-S. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang, "The

- sequence and de novo assembly of the giant panda genome,” *Nature*, vol. 463, no. 7279, pp. 311-317, Jan. 2010.
- [29] N. Rabbee and T. P. Speed, “A genotype calling algorithm for affymetrix SNP arrays,” *Bioinformatics*, vol. 22, no. 1, pp. 7-12, Jan. 2006.
 - [30] Fisher R. A., “The Correlation Between Relatives on the Supposition of Mendelian Inheritance,” *Philosophical Transactions of the Royal Society of Edinburgh*, vol. 52, pp. 399-433, 1918.
 - [31] C. R. Henderson, *Applications of linear models in animal breeding*. [Guelph, Ont.]: University of Guelph, 1984.
 - [32] B. Ewing and P. Green, “Analysis of expressed sequence tags indicates 35,000 human genes,” *Nat. Genet.*, vol. 25, no. 2, pp. 232-234, Jun. 2000.
 - [33] A. E. Shrimpton and A. Robertson, “The Isolation of Polygenic Factors Controlling Bristle Score in *Drosophila Melanogaster*. I. Allocation of Third Chromosome Sternopleural Bristle Effects to Chromosome Sections,” *Genetics*, vol. 118, no. 3, pp. 437-443, Mar. 1988.
 - [34] B. Hayes and M. E. Goddard, “The distribution of the effects of genes affecting quantitative traits in livestock,” *Genet. Sel. Evol.*, vol. 33, no. 3, pp. 209-229, Jun. 2001.
 - [35] M. F. Rothschild, R. G. Larson, C. Jacobson, and P. Pearson, “Pvu II polymorphisms at the porcine oestrogen receptor locus (ESR),” *Animal Genetics*, vol. 22, no. 5, pp. 448-448, Oct. 1991.
 - [36] L. Andersson and M. Georges, “Domestic-animal genomics: deciphering the genetics of complex traits,” *Nat. Rev. Genet.*, vol. 5, no. 3, pp. 202-212, Mar. 2004.
 - [37] A. Darvasi and M. Soller, “A Simple Method to Calculate Resolving Power and Confidence Interval of QTL Map Location,” *Behavior Genetics*, vol. 27, no. 2, pp. 125-132, 1997.
 - [38] W. G. Hill, “Estimation of Effective Population Size from Data on Linkage Disequilibrium,” *Genetics Research*, vol. 38, no. 03, pp. 209-216, 1981.
 - [39] W. G. Hill and A. Robertson, “Linkage disequilibrium in finite populations,” *Theoretical and Applied Genetics*, vol. 38, no. 6, pp. 226-231, Jun. 1968.
 - [40] Balding D.J., “A tutorial on statistical methods for population association studies,” *Nature Reviews Genetics*, vol. 7(10), 781-791., 2006.
 - [41] E. M. Prinzenberg, I. Krause, and G. Erhardt, “SSCP analysis at the bovine CSN3 locus discriminates six alleles corresponding to known protein variants (A, B, C, E, F, G) and three new DNA polymorphisms (H, I, A1),” *Anim. Biotechnol.*, vol. 10, no. 1-2, pp. 49-62, 1999.
 - [42] G. Rincón, A. Islas-Trejo, J. Casellas, Y. Ronin, M. Soller, E. Lipkin, and J. F. Medrano, “Fine mapping and association analysis of a quantitative trait locus for milk production traits on *Bos taurus* autosome 4,” *Journal of Dairy Science*, vol. 92, no. 2, pp. 758-764, Feb. 2009.
 - [43] Moiola B., Contarini G., Avalli A., Catillo G., Orru L, De Matteis G, Masoero G., and Napolitano F., “Short communication: Effect of stearoyl-CoA desaturase polymorphism on fatty acid composition of milk,” *Journal of Dairy Science*, vol. 90: 3553-3558, 2007.
 - [44] Taniguchi M., Utsugi T., Oyama K., Mannen H., Kobayashi M., Tanabe Y., Ogino A., and Tsuji S., “Genotype of stearoyl-CoA desaturase is associated with fatty acid composition in Japanese Black cattle,” *Mammalian Genome*, vol. 14: 142-148., 2005.
 - [45] B. Grisart, W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, “Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation

- in the bovine DGAT1 gene with major effect on milk yield and composition,” *Genome Res.*, vol. 12, no. 2, pp. 222-231, Feb. 2002.
- [46] Crisà A., Marchitelli C., Pariset L., Contarini G., Signorelli F., Napolitano F., Catillo G., Valentini A., and Moiola B., “Exploring polymorphisms and effects of candidate genes on milk fat quality in dairy sheep,” *Journal of Dairy Science*, vol. 93(8):3834-3845, 2010.
 - [47] McLeod I. M., Hayes B. J., Savin K. W., Chamberlain A. J., McPartlan H. C., and Goddard M. E., “Power of a genome scan to detect and locate QTL in cattle using dense single nucleotide polymorphisms,” *Journal of Animal Breeding and Genetics*, vol. 127:133-142, 2010.
 - [48] Hartl D. L., *A Primer of Population Genetics (2nd edn.)*. Sinauer Associates, Sunderland, 1988.
 - [49] M. H. Gallardo, C. A. González, and I. Cebrián, “Molecular cytogenetics and allotetraploidy in the red vizcacha rat, *Tympanoctomys barrerae* (Rodentia, Octodontidae),” *Genomics*, vol. 88, no. 2, pp. 214-221, Aug. 2006.
 - [50] G. H. Hardy, “Mendelian Proportions in a Mixed Population,” *Science*, vol. 28, no. 706, pp. 49-50, Jul. 1908.
 - [51] W. Weinberg, “Über den Nachweis der Vererbung beim Menschen,” *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, no. 368-382, 1908.
 - [52] Wright S., “The genetical structure of populations,” *Annals of Eugenics*, vol. 15, 323-354, 1951.
 - [53] Nei M. and Chesser R.K., “Estimation of fixation indices and gene diversities,” *Annals of Human Genetics*, vol. 47, 253-259, 1983.
 - [54] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559-575, Sep. 2007.
 - [55] J. R. González, L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno, “SNPassoc: an R package to perform whole genome association studies,” *Bioinformatics*, vol. 23, no. 5, pp. 654-655, Mar. 2007.
 - [56] Aulchenko Y. S., De Koning D. J., and Haley C., “Genomewide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis,” *Genetics*, vol. 177(1):577-585, 2007.
 - [57] Amin N., van Duijn C. M., and Aulchenko Y. S., “A genomic background based method for association analysis in related individuals,” *PLoS ONE*, vol. 2(12):e1274, 2007.
 - [58] D. Clayton and H.-T. Leung, “An R Package for Analysis of Whole-Genome Association Studies,” *Human Heredity*, vol. 64, no. 1, pp. 45-51, 2007.
 - [59] A. Ziegler and I. R. König, *A statistical approach to genetic epidemiology : concepts and applications*. Weinheim: Wiley-VCH, 2006.
 - [60] L. Hosking, S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and C.-F. Xu, “Detection of genotyping errors by Hardy-Weinberg equilibrium testing,” *Eur J Hum Genet*, vol. 12, no. 5, pp. 395-399, Feb. 2004.
 - [61] J. K. Wittke-Thompson, A. Pluzhnikov, and N. J. Cox, “Rational inferences about departures from Hardy-Weinberg equilibrium,” *Am. J. Hum. Genet.*, vol. 76, no. 6, pp. 967-986, Jun. 2005.
 - [62] Astle W. and Balding D. J., “Population structure and cryptic relatedness in genetic association studies,” *Statistical Science*, vol. 24: 451-471, 2009.
 - [63] Pryce, J.E., Hayes, B.J., Bolormaa, S., Goddard, and M.E., “Polymorphic regions affecting human height also control stature in cattle,” *Genetics*, vol. 187, 981-984, 2011.

- [64] E. Boerwinkle, R. Chakraborty, and C. F. Sing, "The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods," *Ann. Hum. Genet.*, vol. 50, no. Pt 2, pp. 181-194, May 1986.
- [65] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler, "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nat Genet*, vol. 38, no. 2, pp. 203-208, Feb. 2006.
- [66] Chen W. M. and Abecasis G. R., "Family-based association tests for genomewide association scans.," *American Journal of Human Genetics*, vol. 81, 913-926, 2007.
- [67] V. N. Michelizzi, M. V. Dodson, Z. Pan, M. E. J. Amaral, J. J. Michal, D. J. McLean, J. E. Womack, and Z. Jiang, "Water Buffalo Genome Science Comes of Age," *Int J Biol Sci*, vol. 6, no. 4, pp. 333-349, Jun. 2010.
- [68] A. V. Zimin, A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassel, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg, "A whole-genome assembly of the domestic cow, *Bos taurus*," *Genome Biology*, vol. 10, no. 4, p. R42, 2009.
- [69] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes," *Nucleic Acids Res.*, vol. 27, no. 11, pp. 2369-2376, Jun. 1999.
- [70] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biology*, vol. 5, no. 2, p. R12, Jan. 2004.
- [71] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754-1760, Jul. 2009.
- [72] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589-595, Mar. 2010.
- [73] Liu Y., Qin X., Song X-Z.H., Jiang H., Shen Y., Durbin K.J., Lien S., Kent M.P., Sodeland M., Ren Y., Zhang L., Sodergren E., Havlak P., Worley K.C., and Weinstock G.M. Gibbs R.A., "Bos taurus genome assembly.," *BMC Genomics*, vol. 10, 180., 2009.
- [74] Rona, M., Kliger, D., Feldmesser, E., Seroussia, E., Ezrab, E., Weller, and J.I., "Multiple Quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design," *Genetics*, vol. 159: 727-735., 2001.
- [75] Pariset L., "Analisi di geni candidati," in *I Georgofili, Quaderni*, vol. III, .
- [76] M. Rambeaud and G. M. Pighetti, "Impaired Neutrophil Migration Associated with Specific Bovine CXCR2 Genotypes," *Infection and Immunity*, vol. 73, no. 8, pp. 4955 -4959, Aug. 2005.
- [77] M. Rambeaud and G. M. Pighetti, "Differential calcium signaling in dairy cows with specific CXCR1 genotypes potentially related to interleukin-8 receptor functionality," *Immunogenetics*, vol. 59, no. 1, pp. 53-58, Nov. 2006.
- [78] S. M. Viitala, N. F. Schulman, D. J. de Koning, K. Elo, R. Kinos, A. Virta, J. Virta, A. Mäki-Tanila, and J. H. Vilkkki, "Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle," *J. Dairy Sci.*, vol. 86, no. 5, pp. 1828-1836, May 2003.
- [79] Roy R., Ordovas L., Zaragoza P., Romero A., Moreno J., Altarriba J., and Rodellar A., "Association of Polymorphisms in the bovine FASN gene with milk-fat content," *Animal Genetics*, vol. 37: 215-218., 2006.
- [80] Morris, C. A., Cullen N. G., Glass B. C., Hyndman D. L., Manley T. R., Hickey S. M., McEwan J. C., Pitchford W. S., Bottema C. D., and Lee M. A., "Fatty acid synthase effects on bovine adipose fat and milk fat," *Mammalian Genome*, vol. 18: 64-74, 2007.

- [81] Banos G., Wolliams J. A., Woodward B. W., Forbes A. B., Coffey M. P. (2008) Impact of Single Nucleotide Polymorphisms in Leptin, Leptin Receptor, Growth Hormone Receptor, and Diacylglycerol Acyltransferase, "Gene Loci on Milk Production, Feed, and Body Energy Traits of UK Dairy Cows," *Journal of Dairy Science*, vol. 91: 3190-3200.
- [82] N. A. Ganai, H. Bovenhuis, J. A. M. van Arendonk, and M. H. P. W. Visker, "Novel polymorphisms in the bovine β -lactoglobulin gene and their effects on β -lactoglobulin protein concentration in milk," *Animal Genetics*, vol. 40, no. 2, pp. 127-133, Apr. 2009.
- [83] Di Stasio L., Sartore S., and Lisa C., "Polymorphisms of candidate genes for fatty acids composition in the Piedmontese cattle breed," *Italian Journal of Animal Science*, vol. 8(2):214-216, 2009.
- [84] Ogorevc J., Kunej T., Razpet A., and Dvornik P., "Database of cattle candidate genes and genetic markers for milk production and mastitis," *Animal Genetics*, vol. 40(6): 832-851., 2009.
- [85] Caroli, O. Jann, E. Budelli, P. Bolla, S. Jäger, and G. Erhardt, "Genetic polymorphism of goat kappa-casein (CSN3) in different breeds and characterization at DNA level," *Animal Genetics*, vol. 32, no. 4, pp. 226-230, 2001.
- [86] H. J. Benjamini Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, no. 1, pp. 289-300.
- [87] Kontopidis G., Holt C., and Sawyer L., "Invited Review: β -Lactoglobulin: Binding Properties, Structure and Function," *Journal of Dairy Science*, vol. 87: 785-796, 2004.
- [88] Gimpl G. and Fahrenholz F., "The Oxytocin Receptor System: Structure, Function, and Regulation," *Physiological Reviews*, vol. 81(2): 629-683, 2001.
- [89] Nishimori K., Young L. J., Guo Q., Wang Z., Insel T. R., and Matzuk M. M., "Oxytocin is required for nursing but is not essential for parturition or reproductive behavior," *PNAS*, vol. 93:11699-11704., 1996.
- [90] Uemoto Y., Abe T., Tameoka N., Hasebe H., Inoue K., Nakajima H., Shoji N., Kobayashi M., and Kobayashi E., "Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle," *Animal Genetics*, vol. 42, 141-148., 2010.
- [91] Bruford M.W., Bradley D.G., and Luikart G., "DNA Markers reveal the complexity of livestock domestication," *Nature Reviews Genetics*, vol. 4, 900-910., 2003.
- [92] Laloë D., Moazami-Goudarzi K., Lenstra J., Marsan P., Azor P., Baumung R., Bradley D., Bruford M.W., Cañón J., Dolf G., Dunner S., Erhardt G., Hewitt G., Kantanen J., Obexer-Ruff G., Olsaker I., Rodellar C., Valentini A., Wiener P., European Cattle Genetic Diversity Consortium, and Econogene Consortium., "Spatial Trends of Genetic Variation of Domestic Ruminants in Europe," *Diversity*, vol. 2(6), 932-945., 2010.
- [93] Groeneveld L.F., Lenstra J.A., Eding H., Toro M.A., Scherf B., Pilling D., Negrini R., Finlay E.K., Jianlin H., Groeneveld E., and Weigend S. GLOBALDIV Consortium., "Genetic diversity in farm animals - a review," *Animal Genetics*, vol. 41, 6-31., 2010.
- [94] Taberlet P., Coissac E., Pansu J., and Pompanon F., "Conservation genetics of cattle, sheep, and goats," *Comptes Rendus Biologies*, vol. 334(3), 247-254., 2011.
- [95] Stella A., Ajmone-Marsan P., Lazzari B., and Boettcher P., "Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production," *Genetics*, vol. 185, 1451-1461, 2010.

- [96] Felius M., Koolmees P.A., Theunissen B., European Cattle Genetic Diversity Consortium, and Lenstra J.A., "On the breeds of cattle-historic and current classifications," *Diversity*, vol. 3, 660?692., 2011.
- [97] Moioli B., Napolitano F., and Catillo G., "Genetic Diversity between Piedmontese, Maremmana, and Podolica Cattle Breeds," *Journal of Heredity*, vol. 95(3), 250?256., 2004.
- [98] Beja-Pereira A., England P.R., Bradley D.G., Jann O.C., Bertorelle G., Chamberlain A.T., Nunes T.P., Metodiev S., Ferrand N., and Erhardt G., "Gene-culture coevolution between cattle milk protein genes and human lactase genes," *Nature Genetics*, vol. 35, 311?313, 2003.
- [99] Cymbron T., Freeman A.R., Malheiro M.I., Vigne J.D., and Bradley D.G., "Microsatellite diversity suggests different histories for Mediterranean and Northern European cattle populations," *Proceedings of The Royal Societies B*, vol. 272, 1837?1843, 2005.
- [100] Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P., Lenstra J.A., and Ajmone-Marsan P., "Breed assignment of Italian cattle using biallelic AFLP markers," *Animal Genetics*, vol. 38(2), 147?53., 2007.
- [101] Dalvit C., De Marchi M., and Cassandro M., "Genetic traceability of livestock products: A review," *Meat Science*, vol. 77: 437?449., 2007.
- [102] M. Sodhi, M. Mukesh, B. P. Mishra, S. P. S. Ahlawat, B. Prakash, and R. C. Sobti, "Microsatellite analysis of genetic population structure of zebu cattle (*Bos indicus*) breeds from north-western region of India," *Anim. Biotechnol.*, vol. 22, no. 1, pp. 16-29, Jan. 2011.
- [103] D'Andrea M., Pariset L., Matassino D., Valentini A., Lenstra J. A., Maiorano G., and Pilla F., "Genetic characterization and structure of the Italian Podolic cattle breed and its relationship with some major European breeds," *Italian Journal of Animal Science*, vol. 10, e54, 2011.
- [104] M. Gautier, T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.-G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen, "Genetic and Haplotypic Structure in 14 European and African Cattle Breeds," *Genetics*, vol. 177, no. 2, pp. 1059 -1070, Oct. 2007.
- [105] McKay S.D., Schnabel R.D., Murdoch B.M., Matukumalli L.K., Aerts J., Coppetiers W., Crews D., Dias Neto E., Gill C.A., Gao C., Mannen H., Wang Z., Van Tassel C.P., Williams J.L., Taylor J.F., and Moore S.S., "An assessment of population structure in eight breeds of cattle using a whole genome SNP panel," *BMC Genetics*, vol. 9, 37., 2008.
- [106] M. Gautier, D. Laloë, and K. Moazami-Goudarzi, "Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds," *PLoS ONE*, vol. 5, no. 9, p. e13038, 2010.
- [107] J. Lewis, Z. Abas, C. Dadousis, D. Lykidis, P. Paschou, and P. Drineas, "Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs," *PLoS ONE*, vol. 6, no. 4, p. e18007, Apr. 2011.
- [108] T. A. Pearson and T. A. Manolio, "How to interpret a genome-wide association study," *JAMA*, vol. 299, no. 11, pp. 1335-1344, Mar. 2008.
- [109] Wigginton J.E., Cutler D.J., and Abecasis G.R., "A Note on Exact Tests of Hardy-Weinberg Equilibrium," *The American Journal of Human Genetics*, vol. 76, 887?883., 2005.
- [110] Pritchard J., Stephens M., and Donnelly P., "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, 945?959., 2000.
- [111] Hubisz M.J., Falush D., Stephens M., and Pritchard J.K., "Inferring weak population structure with the assistance of sample group information," *Molecular Ecology Resources*, vol. 9, 1322?1332., 2009.

- [112] Vaysse A., Ratnakumar A., Derrien T., Axelsson E., Rosengren Pielberg G., Sigurdsson S., Fall T., Seppälä E.H., Hansen M.S.T., Lawley C.T., Karlsson E.K., The LUPA Consortium, Bannasch D., Vilá C., Lohi H., Galibert F., Fredholm M., Häggström J., Hedhammar Å., André C., Lindblad-Toh K., Hitte C., and Webster M.T., "Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping," *PLoS Genetics*, vol. 7(10), e1002316., 2011.
- [113] Rosenberg N., "DISTRUCT: a program for the graphical display of population structure," *Molecular Ecology Notes*, vol. 4, 137-138., 2004.
- [114] Weir B.S., Cardon L.R., Anderson A.D., Nielsen D.M., and Hill W.G., "Measures of human population structure show heterogeneity among genomic regions," *Genome Research*, vol. 15, 1468-1476., 2005.
- [115] Villa-Angulo R., Matukumalli L.K., Gill C.A., Choi J., Van Tassell C.P., and Grefenstette J.J., "High-resolution haplotype block structure in the cattle genome," *BMC Genetics*, vol. 10, 19., 2009.
- [116] MacHugh D.E., Loftus R.T., and Cunningham P. Bradley D.G., "Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers," *Animal Genetics*, vol. 29, 333-340., 1998.
- [117] C. Diez-Tascón, R. P. Littlejohn, P. A. Almeida, and A. M. Crawford, "Genetic variation within the Merino sheep breed: analysis of closely related populations using microsatellites," *Anim. Genet.*, vol. 31, no. 4, pp. 243-251, Aug. 2000.
- [118] Ciampolini R., Cetica V., Ciani E., Mazzanti E., Fosella X., Marroni F., Biagetti M., Sebastiani C., Papa P., Filippini G., Cianci D., and Presciuttini S., "Statistical analysis of individual assignment tests among four cattle breeds using fifteen STR loci," *Journal of Animal Science*, vol. 84, 11-19, 2006.
- [119] L. Flori, S. Fritz, F. Jaffrézic, M. Boussaha, I. Gut, S. Heath, J.-L. Foulley, and M. Gautier, "The Genome Response to Artificial Selection: A Case Study in Dairy Cattle," *PLoS ONE*, vol. 4, no. 8, p. e595, Aug. 2009.
- [120] Olsen H.G., Lien S., Gautier M., Nilsen H., Roseth A., Berg P.R., Sundsaasen K.K., Svendsen M., and Meuwissen T.H.E., "Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6," *Genetics*, vol. 169, 275-283., 2005.
- [121] Zheng X., Ju Z., Wang J., Li Q., Huang J., Zhang A., Zhong J., and Wang C., "Single nucleotide polymorphisms, haplotypes and combined genotypes of LAP3 gene in bovine and their association with milk production traits," *Molecular Biology Reports*, vol. 38 (6), 4053-4061., 2010.
- [122] Olsen H.G., Meuwissen T.H., Nilsen H., Svendsen M., and Lien S., "Fine mapping of quantitative trait Loci on bovine chromosome 6 affecting calving difficulty," *Journal of Dairy Science*, vol. 91, 4312-4322., 2008.
- [123] S. Qanbari, D. Gianola, B. Hayes, F. Schenkel, S. Miller, S. Moore, G. Thaller, and H. Simianer, "Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle," *BMC Genomics*, vol. 12, no. 1, p. 318, Jun. 2011.
- [124] Gu J., Orr N., Park S.D., Katz L.M., Sulimova G., MacHugh D.E., and Hill E.W., "A Genome Scan for Positive Selection in Thoroughbred Horses," *PlosONE*, vol. 4, e5767., 2009.
- [125] C. H. Andreasen, M. S. Mogensen, K. Borch-Johnsen, A. Sandbæk, T. Lauritzen, K. Almind, L. Hansen, T. Jørgensen, O. Pedersen, and T. Hansen, "Studies of CTNBL1 and FDFT1 variants and measures of obesity: analyses of quantitative traits and case-control studies in 18,014 Danes," *BMC Med Genet*, vol. 10, p. 17, Feb. 2009.

- [126] Mejdaring A., "Dystocia and stillbirth in cattle - A review of causes, relations and implications," *Livestock Production Science*, vol. 11, pp. 143-177.
- [127] Mejdaring A., "Sire evaluation for calving traits by Best Linear Unbiased Prediction and nonlinear methodology," *Z. Tierzuchtg. Zuchtgsbiol.*, vol. 102, pp. 95-105.
- [128] Holmberg and L. Andersson-Eklund, "Quantitative trait loci affecting fertility and calving traits in Swedish dairy cattle.," *Journal of Dairy Science*, vol. 89, no. 9, pp. 3664-3671, 2006.
- [129] Druet T., Fritz S., Boussaha M., Ben-Jemaa S., Guillaume F., Derbala D., Zelenika D., Lechner D., Charon C., Boichard D., Gut I.G., Eggen A., and Gautier M., "Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map," *Genetics*, vol. 178, 2227?35., 2008.
- [130] Thomasen J.R., Guldbrandtsen B., Sørensen P., and Thomsen B. Lund M.S., "Quantitative trait loci affecting calving traits in Danish Holstein cattle," *Journal of Dairy Science*, vol. 91, 2098?105., 2008.
- [131] Höglund J.K., Guldbrandtsen B., Su G., Thomsen B. Lund M.S. (2009) Genome scan detects quantitative trait loci affecting female fertility traits in Danish and Swedish Holstein cattle. ;*Journal of Dairy Science*; 92, and 2136-43. Holmberg M. Andersson-Eklund L., "Quantitative trait loci affecting fertility and calving traits in Swedish dairy cattle," *Journal of Dairy Science*, vol. 89, 3664?3671., 2006.
- [132] Brand B., Baes C., Mayer M., Reinsch N., Seidenspinner T., and Thaller G. Kühn Ch., "Quantitative trait loci mapping of calving and conformation traits on Bos taurus autosome 18 in the German Holstein population," *Journal of Dairy Science*, vol. 93, 1205?15, 2010.
- [133] Sahana G. and Guldbrandtsen B. Lund M.S., "Genome-wide association study for calving traits in Danish and Swedish Holstein cattle," *Journal of Dairy Science*, vol. 94, 479?86, 2011.
- [134] Freer B., "Easy calving: not so difficult," *Hereford Breed Journal*, vol. 176?177., 2008.
- [135] Van Tassell C.P. and Wiggans G.R. Misztal I., "Implementation of a sire-maternal grandsire model for evaluation of calving ease in the United States," *Journal of Dairy Science*, vol. 86, 3366?3373., 2003.
- [136] Pausch H., Flisikowski K., Jung S., Emmerling R., Edel C., and Götz Kay-Uwe FriesGenome R., "Wide Association Study Identifies Two Major Loci Affecting Calving Ease and Growth-Related Traits in Cattle," *Genetics*, vol. 187, 289?297., 2011.
- [137] L. Grobet, D. Poncelet, L. J. Royo, B. Brouwers, D. Pirottin, C. Michaux, F. Méniéssier, M. Zanotti, S. Dunner, and M. Georges, "Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle," *Mammalian Genome*, vol. 9, no. 3, pp. 210-213, Mar. 1998.
- [138] Devlin B. and Roeder K., "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997-1004, Dec. 1999.
- [139] Hao K, Di X, and Cawley S., "LdCompare: rapid computation of single- and multiple-marker r^2 ," *Bioinformatics*, vol. 23 (2): 252?254, 2007.
- [140] A. Eberlein, A. Takasuga, K. Setoguchi, R. Pfuhl, K. Flisikowski, R. Fries, N. Klopp, R. Fürbass, R. Weikard, and C. Kühn, "Dissection of genetic factors modulating fetal growth in cattle indicates a substantial role of the non-SMC condensin I complex, subunit G (NCAPG) gene," *Genetics*, vol. 183, no. 3, pp. 951-964, Nov. 2009.
- [141] K. Setoguchi, T. Watanabe, R. Weikard, E. Albrecht, C. Kühn, A. Kinoshita, Y. Sugimoto, and A. Takasuga, "The SNP c.1326T>G in the non-SMC condensin

- I complex, subunit G (NCAPG) gene encoding a p.Ile442Met variant is associated with an increase in body frame size at puberty in cattle,” *Anim. Genet.*, vol. 42, no. 6, pp. 650-655, Dec. 2011.
- [142] Hall, “BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT,” *Nucleic Acids Symposium Series*, vol. 41:95-98., 1999.
- [143] Albera A., Groen A. F., and Carnier P., “Genetic relationships between calving performance and beef production traits in Piemontese cattle,” *Journal of Animal Science*, vol. 82: 3440-3446, 2004.
- [144] Mc Pherron A. C. and Lee S. L., “Double muscling in cattle due to mutations in the myostatin gene,” *PNAS*, vol. 94:12457-12461., 1997.
- [145] Carnier P., Albera A., Dal Zotto R., Groen A.F., and Bona M. Bittante G., “Genetic parameters for direct and maternal calving ability over parities in Piedmontese cattle,” *Journal of Animal Science*, vol. 78, 2532-2539, 2000.
- [146] Gutiérrez-Gil B, Williams JL, Homer D, Burton D, and Haley CS Wiener P., “Search for quantitative trait loci affecting growth and carcass traits in a cross population of beef and dairy cattle,” *Journal of Animal Science*, vol. 87, 24-36., 2009.
- [147] J. Kneeland, C. Li, J. Basarab, W. M. Snelling, B. Benkel, B. Murdoch, C. Hansen, and S. S. Moore, “Identification and fine mapping of quantitative trait loci for growth traits on bovine chromosomes 2, 6, 14, 19, 21, and 23 within one commercial line of *Bos taurus*,” *J. Anim. Sci.*, vol. 82, no. 12, pp. 3405-3414, Dec. 2004.
- [148] C. Schrooten, M. C. A. M. Bink, and H. Bovenhuis, “Whole genome scan to detect chromosomal regions affecting multiple traits in dairy cattle,” *J. Dairy Sci.*, vol. 87, no. 10, pp. 3550-3560, Oct. 2004.
- [149] T. Moore and D. Haig, “Genomic imprinting in mammalian development: a parental tug-of-war,” *Trends in Genetics*, vol. 7, no. 2, pp. 45-49, Feb. 1991.
- [150] L. A. Murphy and K. D. Sarge, “Phosphorylation of CAP-G is required for its chromosomal DNA localization during mitosis,” *Biochem. Biophys. Res. Commun.*, vol. 377, no. 3, pp. 1007-1011, Dec. 2008.
- [151] M. N. Weedon, H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans, M. Mangino, R. M. Freathy, J. R. B. Perry, S. Stevens, A. S. Hall, N. J. Samani, B. Shields, I. Prokopenko, M. Farrall, A. Dominiczak, T. Johnson, S. Bergmann, J. S. Beckmann, P. Vollenweider, D. M. Waterworth, V. Mooser, C. N. A. Palmer, A. D. Morris, W. H. Ouwehand, J. H. Zhao, S. Li, R. J. F. Loos, I. Barroso, P. Deloukas, M. S. Sandhu, E. Wheeler, N. Soranzo, M. Inouye, N. J. Wareham, M. Caulfield, P. B. Munroe, A. T. Hattersley, M. I. McCarthy, and T. M. Frayling, “Genome-wide association analysis identifies 20 loci that influence adult height,” *Nat. Genet.*, vol. 40, no. 5, pp. 575-583, May 2008.
- [152] D. F. Gudbjartsson, G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson, P. Zamanovich, P. Sulem, S. Thorlacius, A. Gylfason, S. Steinberg, A. Helgadóttir, A. Ingason, V. Steinthorsdóttir, E. J. Olafsdóttir, G. H. Olafsdóttir, T. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jorgensen, O. Pedersen, K. K. Aben, J. A. Witjes, D. W. Swinkels, M. den Heijer, B. Franke, A. L. M. Verbeek, D. M. Becker, L. R. Yanek, L. C. Becker, L. Tryggvadóttir, T. Rafnar, J. Gulcher, L. A. Kiemeny, A. Kong, U. Thorsteinsdóttir, and K. Stefansson, “Many sequence variants affecting diversity of adult human height,” *Nat. Genet.*, vol. 40, no. 5, pp. 609-615, May 2008.

Acknowledgements

In this last page I wish to thank the people that have helped me in completing this work (hoping not to damage too badly their reputation including their names here. . .). My first thanks go to Dr. Lorraine Pariset and Prof. Alessio Valentini and with whom I have started learning genomics. I owe to them the greater part of the results presented here. Next, I wish to thank Dr. Giovanni Chillemi and Dr. Nico Sanna at CASPUR for their help and the coordinator Prof. Giorgio Pranterà for his kind support. I gratefully acknowledge Dr. Aleksey Zimin for what I have learned in genome assembly over many conference calls. Thanks go also to Tommaso Biagini for his help in the buffalo assembly. I want to thank all the people at Viterbo: Silvia, Maria, Federica, Silvia, Gabriella (girls first!) and Leonardo, Gianluca and Marco. Finally, my last and most precious thank goes to my women Sofia and Serena.