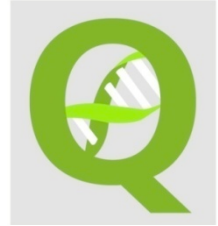




Quantomics



Final Publishable Report

Grant Agreement number: 222664

Project acronym: QUANTOMICS

Project title: From Sequence to Consequence – Tools for the Exploitation of Livestock Genomes

Funding Scheme: Collaborative Project

Period covered: from 1 June 2009 to 30 November 2013

Project's coordinator:

Mr Chris Warkup, Director, Bioscience Network Ltd

Tel: +44 (0)131 651 7334

Fax: +44 (0)131 651 7335

E-mail: Chris.Warkup@biosciencektn.com

Project website address: www.quantomics.eu

EC Project Officers:

Anne-Sophie Lequarré

Jean-Charles Cavitte



This research project has been co-financed by the European Commission, within the 7th Framework Programme, Grant Agreement No. KBBE-2A-222664. The text represents the authors' views and does not necessarily represent a position of the Commission who will not be liable for the use made of such information.



Copying text and data and quoting from this publication is permitted provided with complete and correct acknowledgement of sources.

Table of Contents

Preamble	3
Executive summary.....	4
Introduction.....	5
The project: schematic overview	6
Quantomics work packages: key activities and achievements.....	7
Comparative Genomics.....	8
Bioinformatics Pipeline	8
Allele Discovery by Resequencing.....	9
Validation of Major Alleles	10
Molecular Tools Development & Validation	10
Genome-Wide Selection Tools.....	11
Application of Tools in Commercial Populations on Health Traits	12
Integrated Genomic Management Tools.....	12
Dissemination of Results	13
Facts and figures.....	14
Quantomics: the consortium.....	15
People: contact details	16
Project outcomes: impact.....	17
Bioinformatics: information is power	19
The truth is in the genes: the mastitis story	20
Breeding programmes: where it all comes down to.....	21
Benefits to the livestock sector and society	23
Guidance for the industry	24
Robustness and sensitivities of the evaluation tools.....	27
Opportunities for innovation using Quantomics evaluation tools.....	29
Ethical Considerations.....	31
Glossary.....	33

Preamble

The sequencing of the genomes of the major livestock species has placed animal agriculture on the threshold of a new era. In 2009 the Quantomics Consortium set out to develop tools and methodologies that would overcome the difficulties that existed at the time in making best use of the genomic data.

Animal genomics-related sciences and technologies are developing rapidly. I am proud to have worked with the Quantomics team who showed great dedication to fulfil the project's objectives and who were very flexible in adapting and re-designing the project so we could make best use of the very latest technologies and resources for high-throughput genomics and computational biology.

The whole team is also grateful to the support from our European Commission (EC) Project Officers, and the contributions from third parties and other collaborative projects, which leveraged the financial investment and extended scientific excellence beyond the original project boundaries.

The Project was developed in our aspiration to contribute to the enhancement of animal welfare, health and robustness, thereby helping bring European farmers a step closer to breeding healthier, more productive and higher quality livestock, and supporting the competitiveness and sustainability of the livestock sector.

Toine Roozen
Quantomics Operations Manager

Executive summary

Quantomics, a large collaborative research project co-funded by the European Commission (7th Framework Programme), set out in 2009 to:

- Provide the tools to combine information from reference genome sequences, genetic mapping, comparative genomics, genome annotation, functional studies and dense sequence data from the latest hyper-parallel re-sequencing technologies to identify rapidly the causative deoxyribonucleic acid (DNA) variation underlying important traits in livestock species.
- Provide the methodologies and software tools to exploit effectively and efficiently genomic information in sustainable breeding programmes.
- Determine the benefits and constraints relating to the effective application of these tools to health and welfare traits in chickens and cattle - hence providing information across contrasting breeding schemes that span the breadth of livestock improvement.
- Provide a framework for moving from genome sequence to commercial application for all livestock species which can be expanded to include other forms of information such as epigenetic effects when their importance has been established.
- Disseminate results of the project to ensure broad and long-term beneficial impact on European competitiveness and European Union (EU) policy on animal health and welfare and sustainable agriculture.

Quantomics drew on the complementary expertise of 12 scientific and 5 industry partners to develop, enhance and integrate tools that, capitalising on the recent advances in genomic technology, offer applicable solutions and advise to the livestock breeding sector. These tools fall broadly into the following categories:

- 1) Tools to study and understand the function of individual genes in the animal genome.
- 2) Tools to associate the animal genome with their phenotypes.
- 3) Tools to integrate genomic information into successful breeding programmes.

This was a pan-European, multidisciplinary collaboration that brought together expertise in, genome biology, comparative and functional genomics, bioinformatics, quantitative genetics, statistics, epidemiology and animal breeding. The consortium also benefited by access to different animal resource populations as shown below:

- 38 dairy bulls of two breeds (Finnish Ayrshire and Italian Brown Swiss) were sequenced.
- 1,800 dairy cattle of four breeds (Finnish Ayrshire, Italian Brown Swiss and Valdostana Red Pied, and Danish Red) were genotyped.
- 2,500 chickens were genotyped.

Chickens, cattle, pigs, sheep, turkey and duck were addressed in the project, but the impact of the project's outcomes do not need to be restricted to these species, as they are applicable to other species as well.

Some key results of Quantomics include:

1. The delivery of new bioinformatics analysis and visualisation tools that can be used by scientists to further study the complexities of the animal genome.
2. The delivery of 10's of millions of genetic variants in cattle and chickens associated with annotations that predict functional effects, which can be used by the livestock breeding industry.
3. The delivery of new molecular genetics and quantitative analysis tools, including a new genomic evaluation procedure, which can be used by the livestock breeding industry.
4. The generation of new genomic markers of variation which will help understand why individual animals differ with regards to important traits.
5. The application of these new tools to important health traits in dairy cattle and broiler chickens.

Introduction

Quantomics is a large collaborative research project exploring “From Sequence to Consequence – Tools for the Exploitation of Livestock Genomes”. It is a 4½-year, €8.14 million project involving 12 leading research groups and 5 businesses and is co-financed by the European Commission’s 7th Framework Programme (FP7). The project began on 1 June 2009 and completed on 30 November 2013.

The sequencing of the genomes of the major livestock species places animal agriculture on the threshold of a new era. The development of genomic tools has provided the ability to map genes associated with welfare, quality and production traits. However, the application of these findings in breeding programmes has been hampered by two factors:

- 1) The problem of finding genetic markers sufficiently closely associated with the causative DNA sequence differences to be fully effective in selective breeding.
- 2) The challenge of identifying the causative DNA polymorphisms themselves to provide an optimally informative and portable tool for selection, breeding and understanding trait biology.

In this project we have selected for development of a set of technologies and tools that are designed to maximise our ability to bridge the gap between trait variation and DNA sequence and hence reduce the two identified bottlenecks.

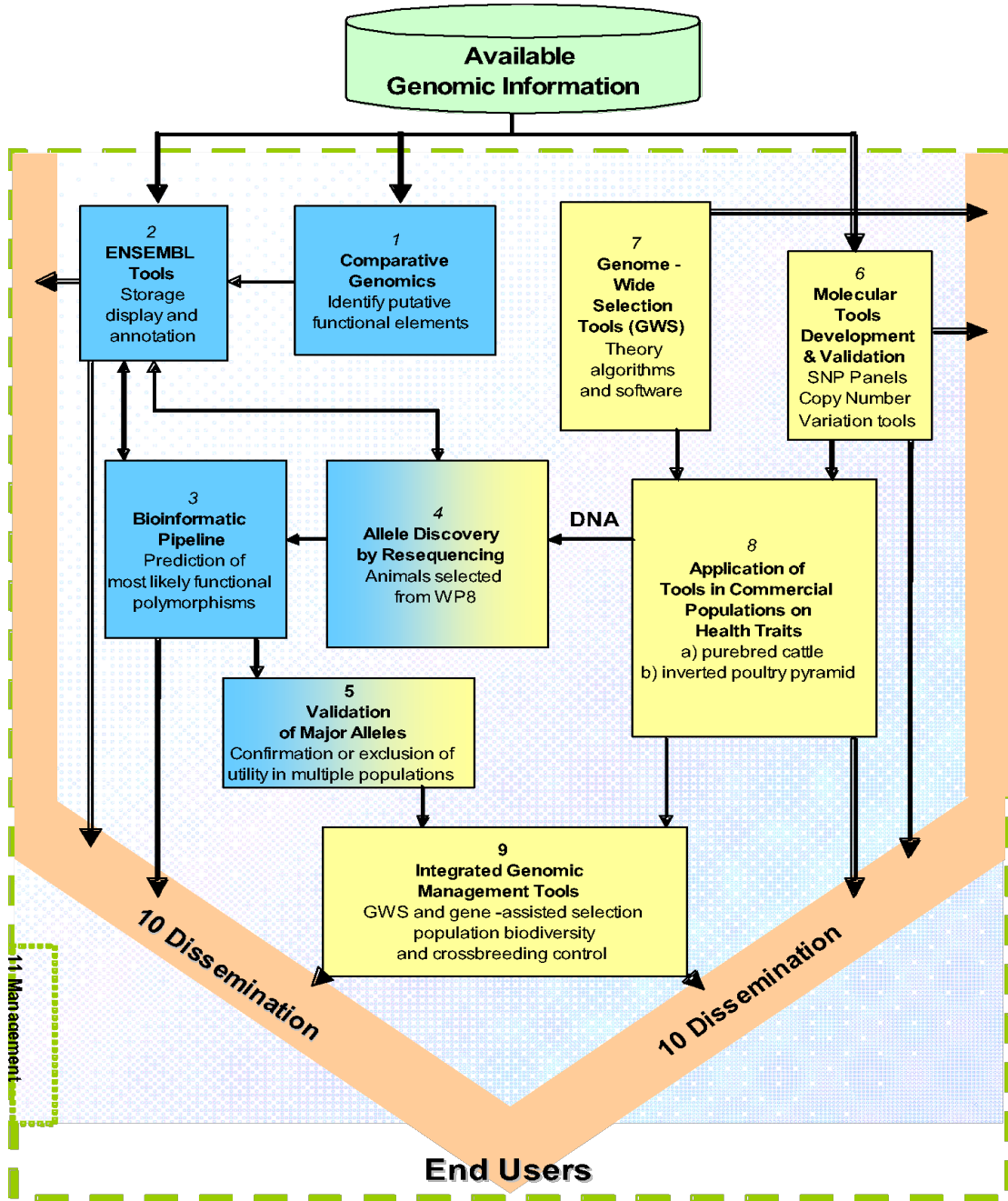
Objectives

Quantomics had from the outset five **high-level Objectives**:

1. *To provide the tools to combine information from reference genome sequence, mapping, comparative genomics, annotation, functional studies and dense sequence data from the latest hyper-parallel re-sequencing technologies to identify rapidly the causative DNA variation underlying important traits in livestock species.*
2. *To provide the methodologies and software tools to exploit effectively and efficiently genomic information in sustainable breeding programmes.*
3. *To determine the benefits and constraints relating to the effective application of these tools to health and welfare traits in chickens and cattle - hence providing information across contrasting breeding schemes that span the breadth of livestock improvement.*
4. *To provide a framework for moving from genome sequence to commercial application for all livestock species which can be expanded to include other forms of information such as epigenetic effects when their importance has been established.*
5. *To disseminate results of the project to ensure broad and long-term beneficial impact on European competitiveness and EU policy on animal health and welfare and sustainable agriculture.*

Quantomics develops new tools for application in the project and by industry, extends these tools to deal with new types of genomic information, such as Copy Number Variations (CNV) or epigenetic information, and provides new tools for the management of biodiversity using molecular DNA information. These tools are expected to have wide application in all farmed species.

The project: schematic overview



Quantomics



Quantomics work packages: key activities and achievements

The key Quantomics activities and achievements have been summarised below, by relevant Work Package (WP).

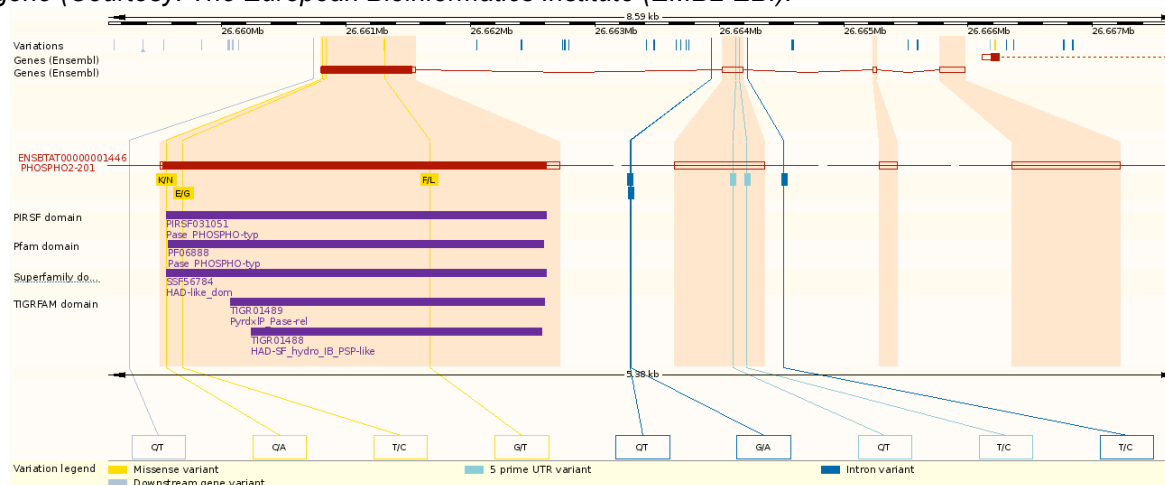
Providing public access to Quantomics Results (WP2)

To ensure the longevity and scientific impact of data and tools produced by Quantomics, it is crucial to make it easily accessible to both inside and outside researchers. The teams involved in this Work Package have focused on the documentation and display of this data and tools, through various means. The WP kept track of the datasets produced by the different Quantomics partners, and made this information available through the Consortium portal. More importantly, the majority of the WP's effort has gone into making Quantomics results publicly available, leveraging well-established resources such as the Ensembl genome browser¹.

The project has provided the tools within Ensembl to store, display and annotate data and present them for exploitation within the rest of the project and for immediate use by outside users. New genomes relevant to the project have been added to or updated in Ensembl.

Quantomics has been successful in developing a number of genomic tools. Some of these, such as the Variant Effect Predictor2, were progressively integrated since Ensembl release 57 (August 2010), whereas others are stand-alone products. The Variant Effect Predictor continues to be in active development, with regular updates (most recently with Ensembl release 75, February 2014). We included many updates over the course of the grant including the integration of farm animal SIFT³ data (since Ensembl release 71, April 2013). Quantomics furthermore developed and deployed new visualisation tools for variation data in Ensembl. Changes in the variation displays include summaries and views that integrate different data types. For instance, the genomic context panel provides information on other genomic features found at the location of a particular variation. Another example is the phenotype view, offering an overview of all the variations associated with a particular phenotype.

Figure 1: With thanks to Quantomics, a researcher can now focus on a gene's variants with the Gene Variation view, which displays a schematic representation of the genes' exons, the protein domains they contain, and the known variants on that gene, with an annotation of the consequences of that gene (Courtesy: The European Bioinformatics Institute (EMBL-EBI)).



¹ <http://www.ensembl.org>

² http://www.ensembl.org/Homo_sapiens/UserData/UploadVariations

³ SIFT: Sorting Intolerant From Tolerant.

Comparative Genomics

The “***Comparative Genomics***” Work Package (WP1) was dedicated to the **bioinformatics integration of data suitable for genomic comparison**. It provided the Quantomics consortium with a way to prioritise genomic regions on the basis of their functional relevance.

The WP initially focused on the genome-wide identification of potential functional elements (PFE). Development of a meta-alignment method for the combination of existing multiple genome aligners resulted in the first prototype of a piece of software, **geno-coffee (now named Robusta)**⁴, which enables combining the output of several aligners into one unique model.

The development of new tools for comparative analysis has also included the computation of multiple genome alignments. Two such alignments have been produced, one including the Turkey⁵, Chicken and Zebra Finch, and the other one adding the Sheep to the reference mammalian genome. The multiple genome alignments have been used for genomic analysis and for the discovery of long non-coding ribonucleic acids (RNAs). The bird alignments have been used by the Turkey Consortium for genomic annotation. The project also generated new RNA sequence data as the most suitable experimental technique to support accurate genome annotation of putative functional elements.

While the original plan was to use mostly evolutionary conservation, the WP scope was subsequently broadened in order to integrate bovine and avian RNA sequencing (RNAseq) data produced by the rest of the consortium. This has led to the report of an improved annotation, based on experimental evidences, and to the identification of a large number of functionally important regions, quite likely to code for long non-coding RNA, which has helped other consortium members in their research for prioritisation of the use of genomic regions when designing breeding strategies.

The farm animal genomes are still far away (completeness- and quality-wise) from i.e. human and mouse genomes. While annotating protein-coding parts of the genome becomes almost a routine procedure, finding functional elements outside coding exons remains a challenge. Hence, multiple, approaches are being used to discover different, but often overlapping sets of sequences, starting from the conserved ones among several species to still poorly characterized long non-coding RNAs. Hence, we used RNAseq mapping and gene calling, as well as multiple genome alignments and mapping functional elements from one species to several others. Thanks to these methods, we significantly improved annotation of putative functional elements for several species of interest.

Bioinformatics Pipeline

The “***Bioinformatics Pipeline***” Work Package develops tools with application in all species. First a fast and sequencing-error resistant sequence-reads mapping tool, **Segemehl**⁶, has been developed, with an aim to find optimally scoring local alignments of a read and the reference genome: it is based on computing inexact seeds of variable length and can handle insertions, deletions (indels; gaps), and mismatches. On real data, this approach has outperformed other read mapping methods. Beyond solving standard mapping tasks, **Segemehl** is particularly suitable for the discovery of spliced, circularized, trans-spliced and other atypical transcripts.

⁴ <http://www.tcoffee.org/Projects/robusta/index.html>

⁵ <http://www.plosbiology.org/article/info:doi%2F10.1371%2Fjournal.pbio.1000475>

⁶ <http://www.bioinf.uni-leipzig.de/Software/segemehl/>

The prediction of polymorphism effects in non-protein-coding RNAs is a young topic. With **RNAsnp**⁷ the first toolkit of this type has been developed, and made available as both a stand-alone implementation for download and as web service.

The tools and methods that deal with the bioinformatics pipeline are integrated in a large, modular genome alignment and RNA annotation pipeline named **rnannotator**.

It aggregates 6 tools for *de-novo* annotation and uses 13 databases for homology annotation, greatly mainstreaming the annotation process, especially when new genome assemblies are published.

Following the construction of this automated pipeline for functional annotation and prioritisation of cattle and chicken polymorphisms were mapped to annotation elements generated within the project or originating from public sources. Aggregating information from a diverse set of tools with a simple scoring scheme allowed prioritisation of cattle and chicken single nucleotide polymorphisms (SNPs). Top-scoring SNPs were used for further validation studies in Quantomics.

Allele Discovery by Resequencing

The “*Allele Discovery by Resequencing*” Work Package (WP4) aimed to obtain the necessary sequence information to identify genetic variation present in the functional elements within the relevant regions of the genomes of cattle and chicken.

The progress in sequencing technology in recent years has made it feasible/affordable to perform whole genome resequencing – rather than trying to sequence specific genome regions of interest. This approach will generate more information on genetic variants (identifying both Copy Number Variants and genetic variants outside the targeted regions) than enrichment procedures of specific genome regions. This approach was therefore adopted by Quantomics. Resequencing of individuals commenced in February 2011. Contact with other projects and organisations, including the United States Department of Agriculture (USDA), was made in order to make best use of data on individuals that have already been sequenced – avoiding duplication of sequencing animals.

Chicken sequencing of pooled DNAs (10 animals per pool) from 24 lines has been accomplished by the University of Edinburgh in collaboration with external partners. Cattle sequencing has also been accomplished and 18 bulls from the Finnish Ayrshire breed and 20 Italian Brown Swiss bulls have been re-sequenced up to more than 20-fold coverage of the genome. Sequence data enabled analysis of the genetic variation influencing different traits in the two species (WP6, WP8).

The main task of this WP was to identify SNPs in the chromosomal regions identified in WP8 as being associated with resistance to avian pathogenic *Escherichia coli* (APEC) in chickens and mastitis in cattle. A total of 443,802 SNPs in 8 chicken chromosomal regions; 327,036 SNPs in Finnish Ayrshire and 299,733 SNPs in Italian Brown Swiss (total 12 chromosomal regions in cattle) were identified for further prioritisation within the project.

The 600k SNPs on the generated chicken Affymetrix SNP chip and a further 15 million high quality SNPs at the end of 2014, 9,816,592 SNPs from Finnish Ayrshire, and 7,515,299 SNPs from Italian Brown Swiss have been deposited in the Single Nucleotide Polymorphism Database (dbSNP). Ensembl extracts the information from dbSNP and incorporates it in the publicly available Ensembl Variation database.

Low coverage genome sequence of two galliform bird species, common quail (*Coturnix coturnix*) and grey partridge (*Perdix perdix*), was generated enabling comparative genome

⁷ <http://rth.dk/resources/rnasnp/>

analysis with other avian genomes such as chicken and turkey. This was done using an Illumina GAI sequencer obtaining 9x and 10x coverage of the two species, respectively. This data, which is available in the sequence read archive at the European Bioinformatics Institute (EMBL-EBI), also corroborates identification of conserved regions and potential regulatory regions in the genomes of chicken and turkey.

Validation of Major Alleles

Samples from two independent cattle populations (Danish Red and Valdostana Red Pied), a new sample of Finnish Ayrshire and a commercial crossbred chicken population were used for the validation of Genome-Wide Association Study (GWAS) results (from WP8). **In total 14 regions in cattle and 10 regions in chicken with identified effects on the target traits (resistance to mastitis in cattle and avian pathogenic *Escherichia coli* (APEC) in chicken) were chosen for the search of functional sequence variations.**

After identification in WP4 and prioritisation according to estimated effect in WP3 (above), a total of 384 cattle and 512 chicken SNPs were included in custom arrays, which were used to genotype validation samples (1145 cattle and 1820 chicken), and test the SNPs for association with mastitis resistance phenotypes in cattle and APEC resistance in poultry. In an *in silico* approach, made possible by recent advances in genotyping and whole genome sequencing, imputed sequences of the cattle QTL regions for 845 Danish Red sires were analyzed for association to mastitis related traits.

In cattle, seven quantitative trait loci (QTL) regions (QTLR) were confirmed by association in multiple breeds. Combining the results with previous functional mastitis data⁸ highlighted several candidate SNPs/genes.

In chicken, RNAseq data⁹ was used to select SNPs from the QTLR for validation and to estimate differential expression of candidate genes. Two QTLR were validated by the custom SNP associations, one of the SNPs being very likely a causative variant.

Molecular Tools Development & Validation

Haplotype blocks and CNV are major tools for achieving a better understanding of the inherited basis of genomic variation in useful livestock phenotypes.

The overall aim of the Work Package for “*Molecular Tools Development & Validation*” was to provide high resolution mapping tools for genome-wide association studies (GWAS) and genome-wide selection (GWS), and to annotate the cattle and chicken genomes for features contributing to quantitative trait genetic variation, including (i) location and approximate boundaries of CNV loci and (ii) haplotype block structure.

Different approaches were taken to characterise CNVs in chickens and cattle: in cattle extensive whole-genome deep sequencing of two breeds (Finnish Ayrshire and Brown Swiss) was used, whereas in chicken Comparative Genomic Hybridization (CGH) was performed on a Nimblegen platform. The Nimblegen platform contains 385,000 oligonucleotides mapped to the assembly of the chicken genome (version 2.1 also known as Galgal3) and cross-referenced to the Ensembl database (release 70). In both species these results were compared to CNV predicted from analysis of dense SNP data.

⁸ Functional mastitis data: obtained from Holstein cattle, part-funded through the EC FP6 project SABRE (contract no. 016250, www.sabre-eu.eu)

⁹ RNAseq data: obtained through collaboration with Iowa State University (ISU), from an APEC challenge study done at ISU www.ans.iastate.edu/faculty/index.php?id=sjlamont

In chicken, 2,260 CNV were identified based on screening of 25 unrelated individuals using the standard Nimblegen 385K probe and pooling results from the published literature. The Red Jungle Fowl was used as a reference to allow mapping of the CNV to the chicken genome reference sequence. Location of these CNV was then used to design a 135K array, which includes 108,045 probes spaced at 2675 base pairs (bp) for CNV detection.

In the chicken populations investigated we found that linkage disequilibrium (LD) between CNV loci and adjacent or embedded SNPs was essentially zero. As the polymorphic information content (PIC) of the CNV was universally low, one would not expect this information to materially affect ability to implement GWAS or GWS. Further work at a higher resolution using Next Generation Sequencing (NGS) approaches may be needed to define CNV smaller than detected using CGH arrays and with higher PIC.

CNV in cattle were mapped using the Illumina 50K and the 777K SNP arrays, in the Italian Brown Swiss and the Finnish Ayrshire cattle. Additionally, whole genome sequence data are used (WP4) to map CNV in these two populations.

In cattle we described a Haplotype block map of SNP/CNP¹⁰ loci. From the validation of 10 CNV regions by NGS in cattle we concluded: (1) NGS-based CNV predictions confirm the candidate regions as well as the HD chip-based results; (2) CNV discovery based on high sequence coverage of about 20X outperforms the (HD) chip-based method (3) the size of the CNVs as predicted by CNVnator show good concordance with the candidates from the representative set of CNVs.

Genome-Wide Selection Tools

At the start of the project the most efficient methods to deal with large-scale applications of GWS were not clear and fast/user-friendly software tools for large populations were not available; while GWS approaches incorporating CNV or epigenetic effects were yet to be developed. **The “Genome-Wide Selection Tools” Work Package sought to deliver genome-wide selection tools for use by the Quantomics project and the industry.** For reasons of computational efficiency it has been determined that the method to deal with large-scale GWS should be iterative and not involve Monte Carlo Markov Chain sampling. The method should also account for the fact that many SNPs will have no or very small effects, since otherwise the method would not be able to make sufficient use of future very dense SNP panels and/or sequence data.

Software has been developed that combines the use of pedigree and marker based relationships, unifying these two sources of relationships into a single-step evaluation in which one combined relationship matrix is produced. This used a novel approach that resulted in unbiased results, overcoming problems associated with previous attempts.

Work on Genome Wide Selection Tools was then directed towards the development of flexible algorithms and software for the estimation of genome-wide breeding values (GEBV) of animals. The theoretical concept allowed specific important polymorphisms to have a greater impact than others on the final GEBV. The algorithm was based on linkage analysis to connect the widely available pedigree data with genotypes, when the latter is not available on all animals.

The software for the estimation of GEBVs was subsequently adapted to incorporate 1) GEBV with imprinting; 2) prior information on SNPs; 3) estimation of GEBV in crossbreeding schemes; and 4) CNV data. Early versions of the software were tested using simulated data provided by WP9. In collaboration with WP8, the final version of the software was tested and GEBV performed in large-scale practical data sets by the industrial partners Associazione Nazionale Allevatori Bovini della Razza Bruna (“ANARB”, cattle) and Landbruk og Fødevarer

¹⁰ SNP: Single Nucleotide Polymorphism. CNP: Copy-number polymorphism

("LF", pigs), and was found to be suitable for large-scale computation of individual polymorphism effects and GEBV, thus enabling genomic selection.

Application of Tools in Commercial Populations on Health Traits

Work Package 8 dealt with the "Application of Tools in Commercial Populations on Health Traits" and was responsible for; (i) providing the necessary practical data to be obtained from the commercial study populations associated with the project and, (ii) the application and evaluation of the new molecular and statistical tools developed by other WPs to this data. Early in the project the main cattle and poultry mapping populations were sampled and genotyped (Illumina Bovine 54K; Illumina Custom 42K for poultry). Genetic and phenotypic data on 1244 Italian Brown Swiss bulls and more than 1500 Finnish Ayrshire bulls are available to the project together with their phenotypes.

Poultry 'Case' and 'Control' samples were collected for APEC in commercial broiler chickens and more than 1000 individuals were genotyped. Estimation of GEBV was performed on data from the Italian Brown Swiss population. GWAS analysis was performed in the Italian Brown Swiss for somatic cell score and in the Finnish Ayrshire for both somatic cell count and clinical mastitis. Following the decision to use a whole genome re-sequencing approach in cattle, the individuals that contribute most to the mapping population were identified for allele discovery resequencing in WP4.

Further genotyping of poultry resulted in 2,339 genotypes (with associated phenotypes). Furthermore, a total of 219 Brown Swiss and 238 Finnish Ayrshire bulls were genotyped with the Illumina Bovine High Density (HD) SNP chip (777K SNPs). Imputation of genotypes to 777K occurred in both populations, after which data was made available to other Quantomics WPs as needed.

In poultry, deconvolution of genotype was performed and the deconvoluted genotypes were used in GWAS. Loci of large effect have been identified using the GWAS approach in the Finnish Ayrshire for Somatic Cell Count (SCC) and for Clinical Mastitis, in the Italian Brown Swiss for SCC (jointly with the Swiss Browns Swiss in collaboration with external data provider Qualitas AG) and in poultry for APEC resistance.

Valdostana Red Pied (VRP) sires were genotyped with Illumina 50K (219 sires) or HD SNP chips (143 sires). Brown Swiss trios (i.e. sire-dam-offspring) were genotyped with Illumina HD SNP chip; The Italian and Swiss Brown-Swiss populations were merged (phenotypes and genotypes). Imputation to HD and GWAS was performed across populations. A GWAS was done in the VRP using selective DNA pooling for milk somatic cell count.

WP7 Software for GEBV estimation was tested on two large populations. The scheduled validation has also been performed on an application population to test validity of the tool, with a comparison of different experimental designs carried out to highlight pros and cons.

Integrated Genomic Management Tools

Through the software developed in WP7 and WP9, we demonstrated that:

- 1) bioinformatics priors can increase the accuracy of prediction of genetic merit;
- 2) managing diversity with genomic selection is improved by genomic measures of diversity;
- 3) genomic diversity can be managed at multiple sites simultaneously using semi-definite programming.

The robustness of these models and their timescale for implementation were assessed.

There are some computational limitations on operation of single step evaluations which will restrict practical application in the short-term but computational algorithms address the

problem more directly. The limitations will also diminish as pedigreed commercial populations are more routinely genotyped, although this will also reduce some of the benefit.

A long-lasting benefit is expected to follow from opening the potential for a new class of evaluations models which more clearly separate effects between identity by descent and correlations among founding gametes. Bioinformatics will inform the refinement of selection objectives in relation to the avoidance of undesirable correlated responses and the management of genotype by environment interactions through allowing them to be better addressed in evaluations, but this will likely be in the medium- to long-term.

The management of diversity using genomic data using genomic relationships across the whole genome is implementable in the short term, but the development to multiple sites will be delayed as genome assemblies become more reliable and as computational power increases. The use of multiple genomic measures will allow, for example, footprints of selection to be shaped by controlling the hitch-hiking about loci that confer a large selective advantage, or control directly the loss of diversity in regions such as the major histocompatibility complex (MHC), or combinations of such objectives in addition to controlling the accumulation of homozygosity across the whole genome.


Dissemination of Results

Quantomics and the European Commission jointly organised a workshop on 26 and 27 September 2011 on **"Stimulating collaborations within European livestock genomics projects"** with involvement of end-users with academic and commercial backgrounds. Results of the project have been presented at renowned international conferences, with some of the most significant results being presented at the *"One-day symposium: Genomic tools and technologies"*, which Quantomics co-organised with the European Association for Animal Production (EAAP) and the Sustainable Solutions for Small Ruminants project ("3SR", EC FP7 Agreement 245140). Partners also provide a 2-day course at the "Winter School" of the Next generation European system for cattle improvement and management project ("Gene2Farm", EC FP7 Agreement 289592) and worked successfully together with the industrial partners in the project. **A Technical Bulletin summarises the project, interpreting the key outcomes for the benefit of the livestock improvement sector.**

By the end of the project, Quantomics scientists had published 60 peer-reviewed articles. Information on additional technical and scientific articles and presentations at international conferences are available via the Quantomics website: www.quantomics.eu.




Facts and figures




**Three livestock species
were studied: cattle,
chickens, pigs**



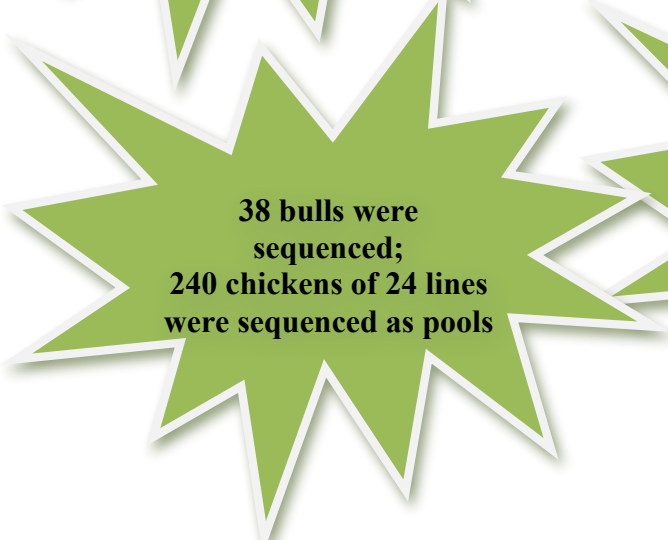
**15 million SNPs defined
and characterised**



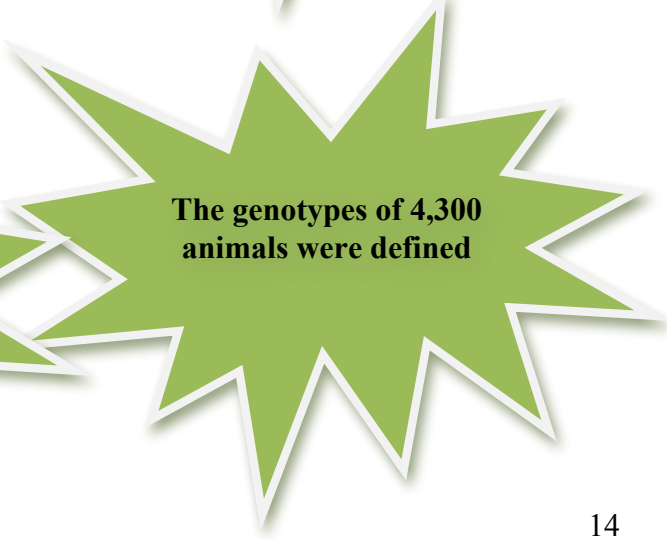
**10 genetic markers
affecting resistance to
APEC were found**



**14 genetic markers
affecting mastitis
resistance were
identified**



**38 bulls were
sequenced;
240 chickens of 24 lines
were sequenced as pools**



**The genotypes of 4,300
animals were defined**

Quantomics: the consortium

Quantomics was established as an inter-disciplinary consortium that brought together complementary expertise from the academia, research and industry:

1. Biosciences KTN, UK – Chris Warkup
2. Aarhus Universitet, DK – Lars-Erik Holm
3. Universitetet for Miljø og Biovitenskap, NO – Theo Meuwissen
4. European Molecular Biology Laboratory, DE – Paul Flicek
5. The University of Edinburgh, UK – John Woolliams
6. Universitat Autònoma de Barcelona, ES – Armand Sanchez
7. Università degli Studi di Milano, IT – Alessandro Bagnato
8. Maa- ja Elintarviketalouden Tutkimuskeskus, FI – Johanna Vilkki
9. Universität Leipzig, DE – Peter Stadler
10. Aviagen Ltd, UK – Kellie Watson
11. The Hebrew University of Jerusalem, IL – Morris Soller
12. Aristotelio Panepistimio Thessalonikis, EL – Georgios Banos
13. Argentix Ltd, UK – Chris Harris
14. Fundacio Privada Centre de Regulacio Genomica, ES – Cedric Notredame
15. Landbrug og Fødevarer, DK – Ingela Velander
16. Commonwealth Scientific and Industrial Research Organisation, AU - Brian Dalrymple
17. Associazione Nazionale Allevatori Bovini della Razza Bruna, IT – Enrico Santus



People: contact details

Coordinator: Chris Warkup

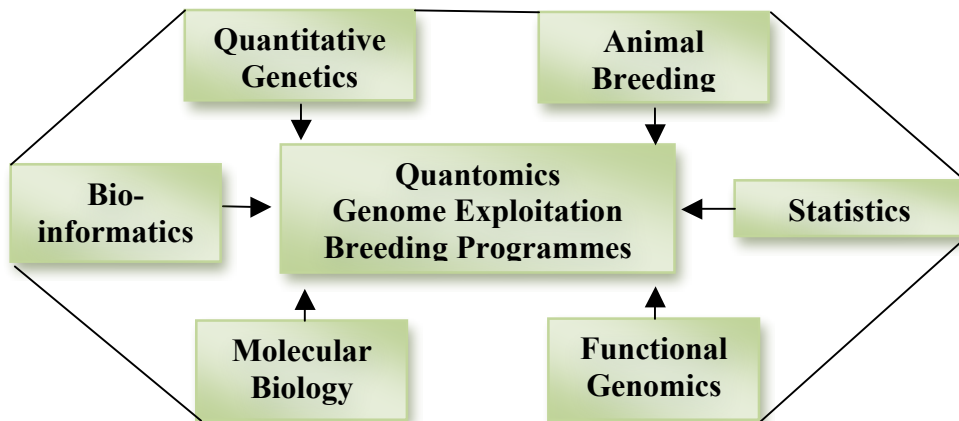
Operations Manager: Toine Roozen

Biosciences KTN, The Roslin Institute, Easter Bush, Midlothian, EH25 9RG, UK

www.innovateuk.org/biosciencesktn info@biosciencektn.com

Work Package Leaders

1. Ensembl, Data storage and annotation: Paul Flicek (Flicek@ebi.ac.uk)
2. Comparative Genomics: Cedric Notredame (Cedric.Notredame@crg.es)
3. Bioinformatics Pipeline: Peter Stadler (Peter.Stadler@bioinf.uni-leipzig.de)
4. Allele Discovery by Resequencing: Lars-Erik Holm (LarsErik.Holm@agrsci.dk)
5. Validation of Major Alleles: Johanna Vilkki (Johanna.Vilkki@mtt.fi)
6. Molecular Tools Development and Validation: David W. Burt (Dave.Burt@roslin.ed.ac.uk)
7. Genome Wide Selection Tools: Theo Meuwissen (Theo.Meuwissen@nmbu.no)
8. Application of Tools in Commercial Populations on Health Traits: Alessandro Bagnato (Alessandro.Bagnato@unimi.it)
9. Integrated Genomic Management Tools: John Woolliams (John.Woolliams@roslin.ed.ac.uk)
10. Dissemination of Results: Toine Roozen (Toine.Roozen@biosciencektn.com)

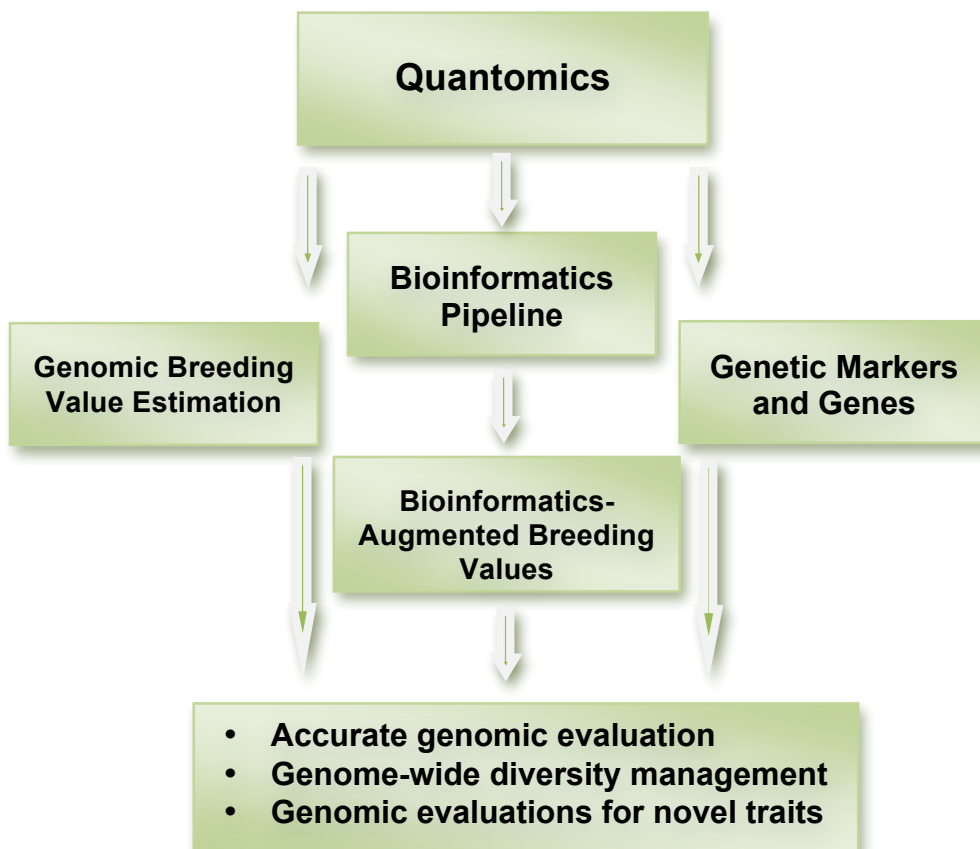


Project outcomes: impact

The sequencing of the genomes of the major livestock species placed animal agriculture on the threshold of a new era. The development of genomic tools has provided the ability to map genes associated with health, welfare, productivity and product quality. However, the application of these findings in breeding programmes has been hampered by two factors:

1. The problem of finding genetic markers sufficiently closely associated with the causative DNA sequence differences to be fully effective in selective breeding.
2. The challenge of identifying the causative DNA polymorphisms themselves to provide an optimally informative and portable tool for selection, breeding and understanding of trait biology.

With the development and implementation of appropriate bioinformatics and analytical tools, Quantomics has managed to identify and validate several hundreds of genetic markers that are putatively linked to individual genes affecting mastitis resistance in cattle and Avian Pathogenic Escherichia coli (APEC) resistance in chickens. Quantomics developed new tools for application in the project and by industry, extended these tools to deal with new types of information, such as Copy Number Variations (CNV) or epigenetic information, and provided new ways for the management of biodiversity using molecular DNA information.

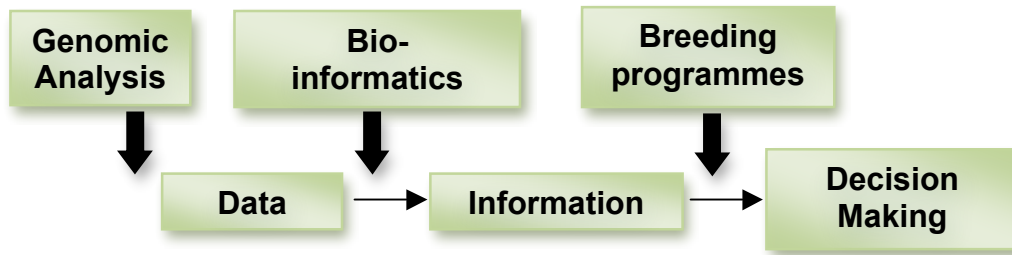


The impact of these activities and outcomes is manifold:

1. Specific genetic markers and genes identified can be used in breeding programmes designed to breed disease resistance into the future generations of dairy cattle and broiler chickens.
2. Methodologies and software developed can be used in further studies of the genomic profile of any livestock species and application to relevant breeding programmes.
3. Results will contribute to the enhancement of animal health and robustness, thereby helping bring European farmers a step closer to breeding healthier, more productive and higher quality livestock, and supporting the competitiveness and sustainability of the livestock sector as a whole.
4. Scientific synergies demonstrated the benefits of inter-disciplinary cooperation and paved the way for future collaborative research activities.
5. Synergies between scientific and industry organisations involved in this project were consolidated and will help the future uptake of research results for the improvement of the livestock sector.
6. More and better links between the genetics and bioinformatics research communities have been established.
7. This has resulted in new tools for annotation of coding and non-coding regions, more powerful annotation of genetic variants, more focused approaches to define causal variants of traits, and ultimately brings us closer to our goal from “Sequence to Consequence”.

Bioinformatics: information is power

The advent of advanced genomic technologies has generated a tremendously large amount of new data pertaining to the genomic profile of individual animals. Bioinformatics tools can be used to manage, organise, visualise and analyse such data, and extract useful information. With appropriate statistical analysis and genetic models, this information is in turn used to inform breeding programmes and facilitate decision-making at farm and/or population level.



Consequently, among the first objectives and activities of the Quantomics project was to develop a pipeline of bioinformatics tools and platforms that could be used within this project but whose utility would also extend across different projects and livestock species.

Some notable outcomes of these activities follow:

1. The bioinformatics annotation (i.e. description) of the livestock genome was enhanced.
2. A bioinformatics pipeline of genome elements (also referred to as markers, variants or polymorphisms) affecting animal health traits was established.
3. Bioinformatics measures that may increase the accuracy of genomic breeding values of individual animals were identified.
4. Useful software was developed, studied or adapted that may:
 - a. Facilitate the study of animal genomes (Pecan, Ortheus, Segemehl).
 - b. Compare genomic profiles across different animal species (Enredo, Robusta): Robusta, in particular, is a novel piece of software developed for the simultaneous comparison of multiple genomes. The utility of these comparisons is in revealing specific parts of a genome that have been protected by evolution because of their importance; they can also reveal parts that have been evolving faster than expected to help species adapt to a new environment. These fast or slow evolving genomic features across various genomes are often functionally important. The main strength of Robusta is its ability to combine the output of multiple genome aligners into one unique model, thereby overcoming a major technical challenge in this area.
 - c. Assess the evolution of animal genomes: Once multiple genome comparisons have been built, it is necessary to quantify the level of natural selection in order to identify the genomic areas most likely to be important. The Genomic Evolutionary Rate Profiling “GERP” measure, deployed in our analysis, is one of the measures able to do this.
 - d. Identify functional features under selection (PipeR): This was another major difficulty in genome analysis of locations without known function. We developed software that can predict the presence of genes in cattle and chicken by similarity to the human genome. These predictions can be useful when studying genomic regions under selection in livestock.
 - e. Predict the role of individual genomic variants (Variant Effect Predictor).

More information on the following website: <http://www.quantomics.eu/tools>

The truth is in the genes: the mastitis story

Using data from the whole animal genome, Quantomics partners identified new genetic variants that are linked with cow resistance to mastitis. The existence of such markers revealed inherent factors that render individual animals resistant to the disease. These markers may be directly incorporated into commercial breeding programmes aiming at enhancing the health profile, and consequently the welfare, of dairy cow populations.

Background and activity

“Mastitis is an inflammatory reaction in the mammary gland due to invasion and colonisation by diverse pathogens”. In other words, the cow udder is sick because of (mostly) bacterial infection. Today, mastitis is the costliest disease affecting dairy cattle, with an estimated annual cost of about €2 billion in Europe alone. Nearly a third of the cow population becomes infected at some time in their life.

In addition to providing good on-farm management and a healthy environment, it has been known for some time that selective breeding can assist in alleviating the mastitis problem – this, because some animals have an inherent capacity to resist infection compared to others. Although well established, this variation had not been well understood at the individual gene level.

The advent of genomic technologies has enabled scientists to study, more closely than before, inherent factors affecting important animal traits.

With assistance from the EC’s FP7 programme, Quantomics set out to develop tools for the exploitation of livestock genome. One of the case-studies conducted to demonstrate the utility of these tools was mastitis in dairy cattle. This was a two-stage study involving four distinct dairy cow populations.

In the first stage, about 500 bulls of the Italian Brown Swiss and the Finnish Ayrshire breed were genotyped with a high-density whole-genome DNA array. Additional genotypes of 2,600 bulls of the same two breeds were made available to the study from other projects. All bulls had previously been progeny tested for milk somatic cell count and clinical mastitis cases of their daughters, thus having an estimate of their genetic merit (estimated breeding value) for mastitis resistance. A Genome-Wide Association Study was performed to match bull genotypes with their estimated breeding values. The analysis identified several markers with a significant effect on the above traits.

In the second stage, these markers were confirmed in two independent populations – the Italian Valdostana Red Pied (nearly 400 bulls) and Red Danish (about 1,000 bulls). Animals were found to have similar phenotypes and genotypes with those of the first stage. This validation exercise lent robustness to the first stage results.

Conclusion

Identified and validated genetic marker results can directly contribute to the on-going efforts to battle dairy cow mastitis in Europe and beyond. Results may also lead to further research on the way individual gene expression affects the manifestation of the disease, thereby promoting our understanding of the udder biology.

The methodologies developed can be applied to other traits and other species.

Breeding programmes: where it all comes down to

Genomics can effectively inform all aspects of breeding programmes, from setting a relevant breeding goal to identifying the genetically superior individuals for selection and proposing optimal mating strategies. Quantomics outcomes contribute to all these steps. In the immediate-term, a relative increase in the genomic evaluation accuracy of the studied traits may be anticipated. In the short- to medium-term, a genomic evaluation procedure including all genotypic, pedigree and phenotypic data in one single step could be in use. In the medium- to longer-term, bioinformatics data will inform the definition of breeding goals and selection objectives, and provide the means for incorporating new traits in genomic evaluations.

Background and activity

A comprehensive breeding programme aiming at enhancing economically and socially important traits of livestock consists of the following steps:

1. Definition of the breeding goal.
2. Data recording for traits of or related with the breeding goal.
3. Estimation of individual animals' breeding values for traits in (2).
4. Selection of breeding animals based on (3).
5. Optimal mating strategies between animals selected in (4).

Variations of this basic scheme have been developed to accommodate differences in the structure of the breeding populations in all species, varying from closely monitored chickens to outbred dairy cattle. The same underlying principle however, has been driving genetic improvement in all livestock species over the past 60 years.

Recent advances in genomics technologies have provided new knowledge and insights that could effectively inform and update the structure of breeding programmes.

The EC FP7 co-funded Quantomics project set out to develop tools for the exploitation of livestock genome. Following the study of the function of individual genes in the animal genome and their association with traits of interest, new tools and methods were developed to integrate this new genomic information into successful breeding programmes.

Simulation studies combined with field data analysis led to a number of relevant outcomes:

1. For the first time, the inclusion of bioinformatics data in the estimation of breeding values was studied. A bioinformatics pipeline developed to identify genomic elements likely to have an impact on breeding goal traits was used to enhance the accuracy of the estimated breeding values of individual animals. Various bioinformatics measures were considered for this matter, with the Genomics Conservation Score showing the greatest promise. Although refinements are still needed, especially with regards to the method applicability to all relevant traits, the proposed approach can benefit breeding programmes in the definition of the breeding goal, and the genetic evaluation and selection steps.
2. An unbiased one-step procedure was developed for the estimation of genomic breeding values that simultaneously incorporates any available genomic, pedigree and phenotypic data. Further computational refinements of this method will enhance its utility in the routine genetic evaluation and selection steps of breeding programmes.
3. New methods to manage genomic diversity and inbreeding across the genome were developed. Genomic relatedness among individuals and optimal contribution theory were used to identify appropriate mating strategies that take into account all sources of new genomic information when predicting future levels of inbreeding. This will benefit

breeding programmes, especially the step relevant to the optimisation of mating strategies between selected individuals.

4. These advances will be particularly beneficial in the introduction into breeding programmes of novel traits of economic importance that are difficult to measure routinely in the field. This can lead to the definition of more comprehensive and relevant breeding goals.

Conclusion

By exploiting recent advances and opportunities in the field of genomics, Quantomics produced new methods for the identification of genetically superior individuals and their use as breeding parents of the next generation. Increases in genetic gain were considered alongside management of biodiversity. Quantomics was able to mobilise resources from partners, produced more avian and bovine genome sequence, millions of SNPs and RNAseq dataset. These developments can effectively inform modern breeding programmes in various livestock species.

Benefits to the livestock sector and society

Tools and outcomes of Quantomics will contribute to the enhancement of animal welfare, health and robustness, thereby helping bring European farmers a step closer to breeding healthier, more productive and higher quality livestock, and supporting the competitiveness and sustainability of the livestock sector.

Quantomics focussed on two very important animal diseases (mastitis in cattle and avian pathogenic *Escherichia coli* (APEC) infection in broiler chickens); therefore, in the first instance, benefits are primarily expected in the form of improved resistance to these conditions.

However, methodologies and tools developed in this project are not species- or trait-specific and can be used for the enhancement of any animal species and trait with socio-economic importance.

Furthermore, improving the health status and welfare of farmed animals will have broader societal ramifications. Healthy animals will produce healthy products that are more readily accepted and sought after by consumers. Enhancing the robustness and welfare of animals will contribute to the sustainability of the sector and its improved image and perception by the society as a whole.

Guidance for the industry

Design of breeding programme

The size and design of the **training (reference) population**, i.e. animals with genotypes and records on which genetic marker effects and, consequently, genomic evaluations are calculated, largely determines the success of genomic breeding programmes. The optimal minimum size will depend on trait and population structure, both demographic and genomic. Relationships among animals should be minimised within the training population and maximised between training and validation animals and selection candidates. In cattle, where the most useful phenotypic data come from progeny-tested bulls, the addition of cows to the training population will be beneficial especially when phenotypes are scarce. The potential impact of this practice would be: (i) the adoption of more sophisticated methods for genomic evaluation, (ii) the expansion of genomic evaluation procedures to include novel, difficult to measure traits, (iii) the development of sound genomic evaluations and comparisons across breeds. More generally, the potential impact is an increase the scope and value of breeding goals and the accuracy with which they are addressed.

Sequencing and genotyping strategies should be based on the genetic contribution of candidate individuals to the reference population. The most informative individuals with the largest contribution should be given priority for **high-density-genotyping and genotyping-by-sequencing**. The contribution should be based on the identify-by-descent relationships including both genomic and pedigree information. The reference population could then be further augmented with the introduction of additional individuals genotyped with **lower-density arrays** at a lower cost (e.g. 65% reduction compared to high-density), which could be handled with appropriate imputation techniques already available. The potential impact of this approach will be on the **improved accuracy of genomic predictions**. The challenge would be the optimisation of the process to identify which exact individuals will be genotyped with each of the different procedures, aiming at maximising the utility of the reference population. Nevertheless the potential impact is to reduce the unit cost of implementing genomic evaluations.

Estimation of genomic breeding value (genomic evaluation)

Unbiased genomic evaluation procedures should be based:

- In the short- to medium-term on the **one-step approach developed and tested in the project**; this method includes linkage analysis information, and corrections for differences in founder population and the finite number of marker loci being used to estimate the genomic relationship matrix; expected benefits compared to previous one-step approaches amount to a nearly **10% increase in accuracy**.
- In the longer term on **bioinformatics-augmented genomic breeding** values that combines bioinformatics data (e.g. the Genomic Conservation Score at each marker/gene location) with genotypes, phenotypes and pedigree. The Quantomics project developed the framework of incorporating such information on genomic evaluations but additional work is required to fully assess the cost-benefit aspects of this concept and render it applicable. The expected impact of this tool will be the development of more accurate, novel genomic evaluations that could potentially **capture a higher proportion of the variance** than the current methods.

As high-density genomic data become available, genomic evaluation methods should be placing **more weight on individual informative markers**. Less informative markers should still receive some lesser weight. Pedigree data should be included to allow the smooth application of the method to genotyped and ungenotyped individuals. The impact of this practice will be an increase in the accuracy of genomic evaluations. The advantage is that actual genes based close to markers with a large effect would duly play a major role in the genomic evaluation. The disadvantage would be that, if marker effects are not being accurately estimated but markers are still given considerable weight, genomic evaluation accuracy will be adversely affected. However, as methods for accurate genomic predictions improve, the latter is expected to become a lesser issue.

When high-density genomic data are used, use of haplotypes does not seem to improve the accuracy of genomic evaluations compared to using individual markers. Therefore, **emphasis should be placed on the accurate estimation of the latter**.

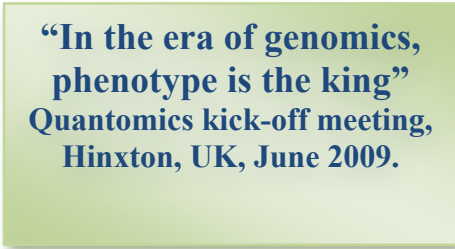
At least **four generations of pedigree** are needed to capture the majority of information in a genomic evaluation using identity-by-descent relationships. The advantage of this practice is that the accuracy of genomic evaluation will increase. Consistent pedigree recording will be required, which may entail costs, but this generally constitutes an integral part of modern breeding programmes.

At the current stage of development, genomic evaluations derived for a certain breed should not be expected to produce any useful prediction of breeding values in another. Possible use of this information in another breed not involved in the design of the DNA array and the calculation of marker effects may severely and adversely affect the genomic evaluations of the breed in question, thereby decreasing the accuracy of genomic predictions to unacceptably low levels.

Mating and inbreeding control

When genomic data is incorporated in the routine genetic evaluations on which selection decisions are based, inbreeding needs to be controlled at the genomic (rather than the pedigree only) level. This implies that relationships between individuals to be mated need to be monitored at the genetic marker level, to ensure that desirable homozygosity is achieved at specific genomic regions harbouring major genes without compromising the overall diversity. The impact of such practice would be that all genomic information is being optimally combined, leading to **maximum genetic progress for given levels of inbreeding allowed or minimal inbreeding for given levels of genetic progress**. The possible difficulty with this procedure is the computational complexity of estimating genomic inbreeding at the marker level including pedigree (i.e. inbreeding based on identity-by-descent). An approximation would be to consider inbreeding based on identity-by-state, which is much easier to compute. However, the exact degree of proximity between the two methods still needs to be determined.

Control of inbreeding in multiple, different regions of the genome and pedigree can be accomplished using **semi-definite programming**. The potential impact of this approach would be that inbreeding will be monitored more accurately and specific regions of interest in the genome (e.g. harbouring major genes) will be observed more closely. However, investment in developing appropriate software will be needed.



**“In the era of genomics,
phenotype is the king”
Quantomics kick-off meeting,
Hinxton, UK, June 2009.**

Robustness and sensitivities of the evaluation tools

Several Quantomics evaluation tools have been developed. Developments were concerned with shaping evaluation tools to better reflect the underlying biology of the traits and to deal better with practical data structures, and to provide more comprehensive options for managing diversity. Here the robustness and sensitivities of these tools are reported.

Single-step evaluations

Single-step evaluations make use of genotyped and ungenotyped animals in a seamless manner.

1. The method developed depends on linkage analysis and this requires genotyped data across a number of generations in order to trace the inheritance from one generation to the next by linkage analysis. Thus the benefits are sensitive to genotyped datasets that are shallow in the sense of few genotyped generations. In such datasets the benefits will be smaller.
2. The inversion of the G matrix is a problem in large practical data sets. A method to avoid the inverse of the G matrix has been developed (Odegard and Meuwissen, 2013). Furthermore the ability to tackle huge datasets is advancing as computing power increases and algorithms to exploit this power are developed, for example through the use of graphical processing units, or the distribution of memory alongside the distribution to parallel processors.
3. When some genotypes are imputed, the linkage analysis based method may be sensitive to imputation errors, as it relies on the genotypes to fit the pedigree structure, whereas many popular imputation methods are based on LD and do not require the imputed genotypes to fit into the genotypes of the family. Note that if no imputation is required the method requires the pedigree but does not require a map! Therefore precise benefits when imputation is required will depend on the quality of the imputation. This will in turn depend on how genotypes have been obtained (by SNP chip or by sequencing and with what coverage), the type of algorithm used (heuristic or Hidden Markov Model) and the software used for implementing the algorithm.
4. The linkage analysis based methods rely mainly on the information from relatives, and thus there is a need for relatives to be included in the training data set. It will also be important that there is regular recruitment of relatives over generations to accumulate more accurate characterisation of the founding genomes.
5. Although this places some limitations on operation, as pedigreed commercial populations are more routinely genotyped there will be an important benefit from developing a new class of models which more clearly separate effects between identity by descent and correlations among founding gametes.

Bayesian tools

Bayesian tools that increase the weight of certain regions based on biology and bioinformatics information:

1. These tools rely on some regions / SNPs being much more important than others. The benefit from this will be greatest when the total number of QTL/genes is substantially lower than the effective number of segments in the population (a measure of the structure of the genome). Therefore will depend not only on the genetic architecture of the trait but also on the population (species and breed within species.) This is because improvements due to the use of bioinformatics data will be sensitive to the relevance of the bioinformatics data for the trait. If the bioinformatics data is not able to discriminate between regions relevant for the trait and other

regions, the accuracy of predictions will improve little. Discrimination between regions will be affected by the extent of LD in the population, in that an additional weight given to variation at one locus will, in effect, give some additional weight to variation neighbouring regions. The search for such precision in bioinformatics data is of utmost importance for the success of evaluation tools that use bioinformatics input.

2. Improvements due to the use of CNV data are sensitive to the genetic variance of the trait explained by the CNV and the extent by which CNV can be predicted by SNPs which are in LD with the CNV. The precise contribution of CNV will become clearer in the short-term but is expected to vary between traits and between breeds. The evidence to date suggests that SNPs loci are able to predict genotypes at CNV loci.
3. Sequence data may improve the Quantomics evaluation tools as here selection can be directly on the causative mutations instead of on LD between the causative mutations and SNPs on the SNP chip. This benefit is theoretical and is as yet unknown, as evaluations with existing sequence are constrained by computational power and existing sequence is likely to be insufficient. However this will change in the medium-term.

Management of diversity

Management of diversity across multiple regions of the genome is achieved using Semi-Definite Programming (SDP).

1. The solution with SDP is robust as it arrives at the solution iteratively whilst moving through feasible solution space, therefore iterations provide closer and closer feasible approximations to the optimum. The optimality at the solution is also guaranteed. This is distinct from other methods which approximate the solution by travelling through infeasible solution space.
2. SDP can cope with large numbers of quadratic and linear constraints.
3. Solutions in the Quantomics problems often show dominance of some constraints over others but this reflects the conditioning of the problem and is not a barrier to use.
4. When using genomic relationships, there is a risk that some procedures for constructing the relationship matrix will result in some negative eigenvalues, which will invalidate SDP. Some procedures for construction can avoid this.
5. Computation of the solution may become a problem when using current technology with large practical datasets. Nevertheless computational algorithms to tackle huge genomic datasets in livestock breeding are being developed; for example using graphical processor units, and the distribution of memory across cores.
6. The use for multiple regions may be delayed until genome assemblies are more reliable, e.g. the MHC in cattle presents sequencing challenges, which requires the continuation of current community efforts.

Opportunities for innovation using Quantomics evaluation tools

It is often hard for the users of technology to understand the stage of development of new technology, in particular when it may have value for the user. This guidance is intended to inform users of the potential impacts and timescales for impact of the technologies developed in Quantomics.

This task involves identifying those indicators and, possibly, contra-indicators of when Quantomics technology can add progress to existing traits of interest, or offer new approaches to improving traits that are currently poorly addressed or not addressed. The final version of the report is provided in Section 4.

Future Perspectives

Implications of Quantomics for implementation in a short to medium term perspective (up to 5 years):

1. *Industry.* Industry want a genomic pipeline for evaluations, a 'sausage machine', which feeds large volumes of genomics, pedigrees, and phenotypes in one end, and in a short time delivers environmentally adjusted, genetically sound predictions of breeding value, preferably multi-trait and preferably unbiased.
2. *Evaluation technology.* The primary driver of accuracy under the control of the breeder is the numerical size of training sets. Model complexity may offer 1.05-fold changes in accuracy and only for a subset of traits.
3. Breeders committed to genomics will increase the size of the training set at a fast rate, partly through imputation, so that any model-induced lack of accuracy (from using a flexible, sound and fast evaluation method c.f. slow and more sophisticated) will be compensated for after a short lag by more data. If this is not true then we must query the adequacies of any of the models and our understanding of the methods!
4. The implication of (1), (2) & (3) is that the types of model that will be fitted in practice are constrained, most probably to be variants of mixed models.
5. *Quantomics.* The project has produced three general breakthroughs for breeders, one associated with the bioinformatics annotation of the genome of the target species for improvement and change, one associated with an unbiased one-step evaluation procedure, and one associated with simultaneous control of multiple measures of diversity. The latter is particularly relevant to managing genomic diversity as Quantomics had shown genomic selection requires genomic control of diversity.
6. Bioinformatic advances in (5) can be incorporated into mixed models and hence can be useful to breeders. However the advances in accuracy are likely to be less than 1.05 fold, particularly if they are intended for use in a 'sausage machine'.
 - Bioinformatic relevance of a SNP for one trait may not be the same for the next trait, and inclusion of trait specific bioinformatics data will need to adapt procedures to multi-trait.
 - Prior to the project it was commonly held that 90+% of the genome was 'junk' therefore annotation would provide the resources for enrichment of information by

focusing on 'active' SNP. The Encyclopedia of DNA Elements (ENCODE¹¹) has made this assumption questionable, perhaps reversed it. This may indicate a large number of SNPs may be active for complex traits.

- However livestock genomes are comparatively coarse and therefore the potential for fine scale weighting of many SNPs is limited.
- BUT the knowledge gained is important for the medium term and beyond if we are going to address correlated responses, and Genetics by Environment (GxE).

7. The one-step procedure mentioned in (5), proposed by Universitetet for Miljø og Biovitenskap (UMB), is a serious way forward and would be implemented if it were less computationally challenging. Overcoming these challenges should be a priority but it would be useful to document these challenges. They are likely to have solutions in the short-term.

8. The opportunity to manage multiple measures of diversity opens up new opportunities to simultaneously manage diversity related to pedigree, to the genome in its entirety, or to specific regions of the genome. This will considerably extend the options for breeders. The use of genomic measures will allow, for example, footprints of selection to be shaped by controlling the hitch-hiking about loci that confer a large selective advantage, or control directly the loss of diversity in regions such as the MHC, or combinations of such objectives in addition to controlling the accumulation of homozygosity across the whole genome. Management of genomic diversity need not have access to a pedigree.

9. *Short-term.* The conclusion of (1) to (8) is that Quantomics technology will not deliver immediate increases in accuracy in the sense of a general pipeline, although some increases in accuracy for particular traits using particular models will be observed.

10. *Short to medium-term.* Unbiased one-step procedures from Quantomics will be in use. Implementation of genomic diversity management is unlikely to be given the same priority by industry and will require some computational refinement.

11. *Medium to long-term.* Bioinformatics will inform the refinement of selection objectives in relation to the avoidance of undesirable correlated responses and the management of genotype by environment interactions through allowing them to be better addressed in evaluations. Genomic diversity should be implemented.

Opportunity

With these perspectives in mind it is relevant for researchers and funding providers to address the refinement of computational algorithms for linkage analysis implementations of one-step and diversity management with multiple constraints.

¹¹ <http://www.genome.gov/encode/>

Ethical Considerations

Ethical considerations have been high on the agenda in the design and implementation of the Quantomics project, and have been an agenda topic at each Quantomics Board Meeting.

Quantomics partners have been (and still are, after the completion of the project) committed to better organise research and to promote the "reduction, refinement and replacement" concept (the "3Rs") in order to find alternatives to animal experimentation.

The Quantomics project collected phenotypes already recorded for cattle, or already recorded or obtained post mortem for poultry; there have been no animal experimentation related to the recording of phenotypes in the project. Most DNA for the cattle studies has been obtained from stored semen samples of existing progeny tested bulls or derived from existing stored DNA samples from national recording programmes or other on-going research within participant organisations.

For poultry, the samples for the case-control study have been derived post-mortem from animals that have been slaughtered in the normal course of food production. By making additional use of traits measurements and samples already approved elsewhere, or obtained post-mortem from normal commercial animals, our approach is fully consistent with the 3Rs - obtaining new knowledge without the need for the use of additional experimental animals.

At one stage in the project, challenge studies were envisaged for the validation of results obtained. Guidelines for an APEC challenge in broilers have been developed at the P3 facility at Universitat Autònoma de Barcelona (UAB), and the APEC challenge experiment and sample collection that was designed for Quantomics had received the approval of the Animal Experimentation Ethics Committee of UAB.

In the end, this experiment was not carried out and ethical considerations were a significant part of the considerations; suitable alternatives to this experiment were found by using data of a challenge study that had taken place in parallel to and outside of the Quantomics project.

Gender Equality

Activities to address gender equality within the project have been put in place by the consortium.

As none of the Consortium partners were seen as not complying with ethics concerning equal opportunities, the Gender Action Plan has been implemented with a 'light touch', without the need for an external expert; During the 1st Periodic Review the progress on Gender Equality Actions had been identified as satisfactory; activities in following periods have been in line with these in the 1st Period:

Gender matters have been an agenda item for all project board meetings, and have been discussed as such. Emphasis hereby has been put on that the implementation will be decentralized, hence to take place within each of the Consortium partners. No issues in relation to gender inequality have been brought to light by any of the partners, or during any of the discussions. Organisations have shown that they address good employment practices, with specific attention to good employment practices for women in science, through other initiatives. To illustrate this, (at least) the following Quantomics Consortium Members are signatories to EC's "The European Charter for Researchers" and "The Code of Conduct for the Recruitment of Researchers":

- Universitat Autònoma de Barcelona (UAB, Spain)
- Centre for Genomaic Regulation (CRG, Spain)
- The Hebrew University of Jerusalem (HUJI, Israel)
- Aristotle University of Thessaloniki (AUTH, Greece)
- Norwegian University of Life Sciences (UMB, Norway)
- The European Bioinformatics Institute (EMBL-EBI, Europe)
- Universities UK, which includes The University of Edinburgh, (UEDIN, UK)

Moreover, The Roslin Institute, University of Edinburgh (UEDIN) received in 2011 a Bronze Athena SWAN award in recognition for its commitment to women's career development.

And the FP7 project "STAGES: Structural Transformation to Achieve Gender Equality in Science" involves the University of Milan (UMIL) and Aarhus University (AU).

Glossary

3SR:	“Sustainable Solutions for Small Ruminants”. An EC co-funded FP7 project
APEC:	Avian Pathogenic <i>Escherichia Coli</i>
bp:	Base Pair, Mb – Megabase, Gb – Gigabase
CGH:	Comparative Genomic Hybridization
CNP:	Copy Number Polymorphism
CNV:	Copy Number Variation
dbSNP:	Single Nucleotide Polymorphism Database
DNA:	Deoxyribonucleic Acid
EAAP:	European Association for Animal Production
EC	European Commission
EMBL-EBI:	The European Bioinformatics Institute (part of European Molecular Biology Laboratory)
EU:	European Union
FP7:	The EC’s 7 th Framework Programme
GBEV:	Genomic Estimated Breeding Value; or Genome Wide Estimated Breeding Value
Gene2Farm:	“Next generation European system for cattle improvement and management project”. An EC co-funded FP7 project
GERP:	Genomic Evolutionary Rate Profiling (GERP)
GWAS:	Genome Wide Association Study
GWS:	Genome Wide Selection
Indel:	DNA mutation (abbreviation of INsertion and DELetion)
LD:	Linkage Disequilibrium
MHC:	Major Histocompatibility Complex
NGS:	Next Generation Sequencing
PFE:	Potential Functional Elements
PIC:	Polymorphic Information Content; measure on how informative a marker is
QTL:	Quantitative Trait Loci
QTLR:	Quantitative Trait Loci Regions
QTS:	Quantitative Trait Site, used in a broad sense for QTN or other polymorphic sequence site having possible functional effects such as an indel or CNP
RNA:	Ribonucleic Acid (RNAs: Ribonucleic Acids)
RNAseq:	Also called "Whole Transcriptome Shotgun Sequencing" ("WTSS"), refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content
SCC:	Somatic Cell Count
SDP:	Semi-Definite Programming
SIFT:	Sorting Intolerant From Tolerant; Tool to predict non-synonymous/missense data
SNP:	Single Nucleotide Polymorphism
TFBS:	Transcription Factor Binding Site
USDA:	United States Department of Agriculture
WP:	Work Package