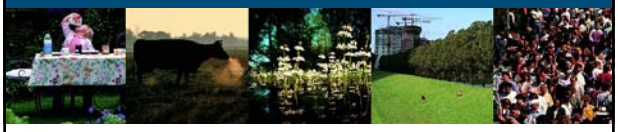
SNP discovery and analysis of selective sweeps using massive parallel short-read sequencing

Martien Groenen



ANIMAL SCIENCES GROUP WAGENINGEN UR

Animal Breeding & Genomics Centre

Research objectives

- Changes in the genome of farm animals and wild life species due to:
  - Speciation
  - Breed formation
  - Selection

SNP chips

Direct sequence comparisons

ANIMAL SCIENCES GROUP WAGENINGEN UR

Animal Breeding & Genomics Centre

SNP discovery: two scenario's

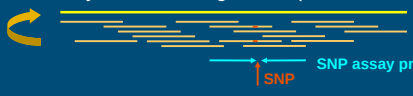
- Reference genome available
  - Pigs
  - Chicken
- No reference genome available
  - Turkey
  - Duck
  - Great tit
  - Tilapia
  - Goat
  - Salmon (planned)

ANIMAL SCIENCES GROUP WAGENINGEN UR

Animal Breeding & Genomics Centre

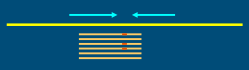
1: Without reference genome: 3 strategies used

(1) Assembly of reference genome (RRLs of 3000 bp fragments)



Turkey
Tilapia
Duck
Great tit

(2) Combination of and Illumina 454 reads



Pig and chicken (missing chromosomes)

(3) Long paired end Illumina reads covering complete fragment

Goat
Duck

ANIMAL SCIENCES GROUP WAGENINGEN UR

Animal Breeding & Genomics Centre

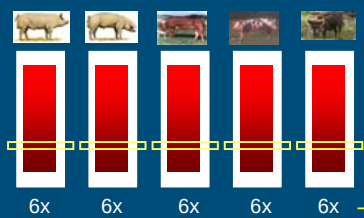
2: Reference genome available

- Pigs
  - Only 100,000 SNPs available before start of project and only a few with minor allele frequency (MAF)
  - 400,000 high quality SNPs + MAF
  - 60K Ilumina iSelect BeadChip
- Chicken
  - 3 million SNPs but no MAF
  - 350,000 high quality SNPs + MAF
  - 60K Ilumina iSelect BeadChip

ANIMAL SCIENCES GROUP WAGENINGEN UR

Animal Breeding & Genomics Centre

1: Reference genome available



- 5 Reduced representation libraries (RRL)
- Restriction digest of DNA

150-250 bp

6x 6x 6x 6x 6x → Variation within breeds

30x → Highly reliable SNP identification and overall MAF

ANIMAL SCIENCES GROUP WAGENINGEN UR

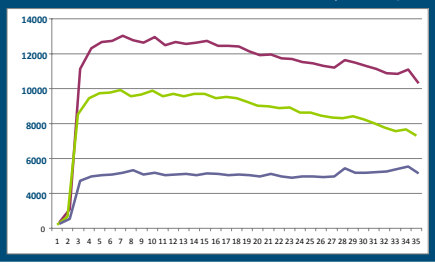
Animal Breeding & Genomics Centre

Align to reference genome and identify SNPs using MAQ

Estimated 10 % of genome covered; Total 400 million 36 bp reads

Distribution of SNPs across GA reads

- Total: 400,000 SNPs identified
  - 333,000 minor allele seen ≥ 3 times
  - 67,000 minor allele seen 2 times + sequence Q minor allele= 40



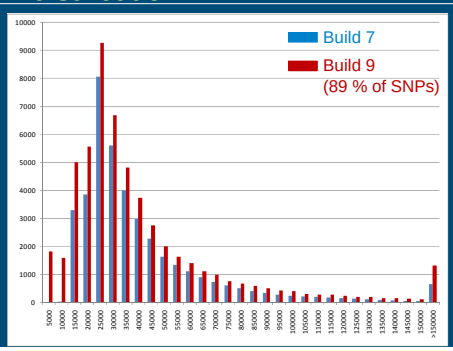
All porcine SNPs currently available

SNP source	Number
HQ Solexa SNPs WU*	333,000
454 sequencing MARC	110,000
7K iSelect chip AU-RI	5,500
dbSNP	17,700
INRA (Sanger seq)	61,700
Cambridge University	14,900
Total No SNPs	543,000
Total Unique SNPs	**510,000**

*Plus additional 58,000 SNPs where minor allele seen twice

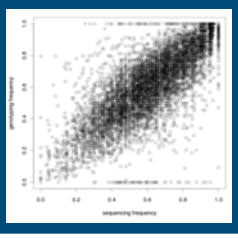
- TOTAL #SNPs submitted for the chip: 72,000
  - 70 % mapped on build 7
  - 7 % predicted
  - 23 % unmapped

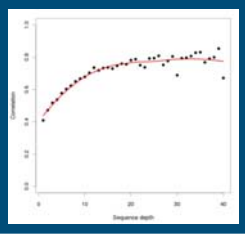
SNP distribution



Porcine 60K BeadChip: Excellent performance

- Final total number of SNPs on the chip: 62,163
- Number of SNPs with MAF>0.05 in at least 1 breed: 59K
- Conversion rate overall: 94.8% (Illumina GA SNPs: 96 %)
- Good correlation between allele count based on GA sequencing and Geotyping



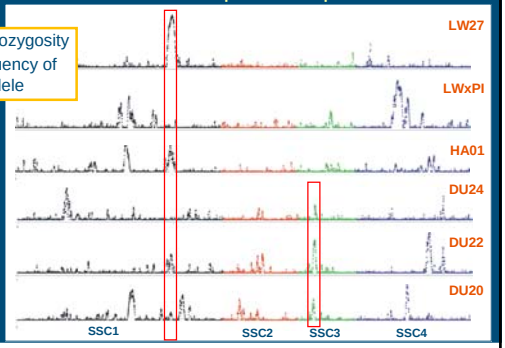


The porcine HapMap project

~1700 Individuals from European and Asian breeds, Wild boar from across Europe and Asia and museum samples

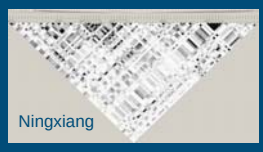
Identification of selective sweeps: Examples on SSC1-4

- Low heterozygosity
- High frequency of derived allele



Use of the 60K chip: European vs Chinese breeds

- Biased towards common variants
- Biased towards SNPs in European breeds
- SNP density of the 60K chip too low for Chinese breeds



Amaral et al. (2008) Genetics 179: 569

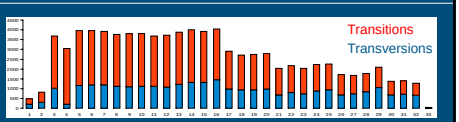
From SNP typing to resequencing

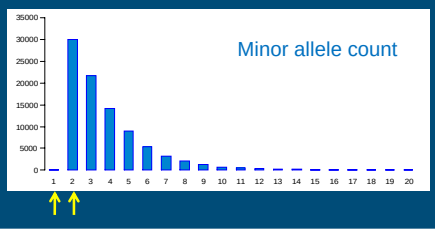
(1) Site frequency spectrum analysis based on RRL sequence data

- RRL cover 5-10 % of the genome
- Coverage differs between breeds
- 4 commercial white breeds + Wild Boar
- Data from pools
- Within 500 Kb windows estimate:
  - Watterson's estimator: $\theta = f(S, n)$.
  - Tajima's estimator: $\pi = f(S, n, freqs.)$.
  - Tajima's D: $D = (\pi - \theta) / sd$.
  - Fst measures population differentiation

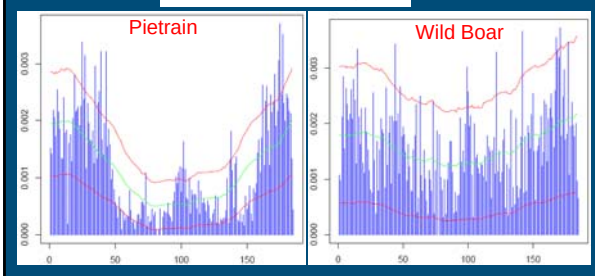
Reanalyse data per breed

SNP distributions





Variation of nucleotide diversity



Suidae tree

~ 1 MY
~ 2 MY
~ 4 MY
~ 10 MY
~ 50 MY

Sus scrofa
Sus barbatus
Sus cebrifons
Sus celebensis
Sus verrucosus
Potomochoerus l.
Potomochoerus
Phacochoerus a
Babyrous
a
Pecari tajacu

Species	Average heter.
Sus scrofa	23 %
Sus barbatus	3 %
Sus celebensis	2 %
Sus cebrifons	0.8 %
Sus verrucosus	0.6 %
Phacochoerus	0.8 %
Potomochoerus l.	0.7 %
Potomochoerus p.	0.9 %
Babyrousa b.	0.9 %

ANIMAL SCIENCES GROUP WAGENINGEN UR
Animal Breeding & Genomics Centre

Sequencing of Sus Verrucosus (Java warty pig)


~2 My

- 8x on Illumina GA
- 200, 500 and 3000 bp libraries

ANIMAL SCIENCES GROUP WAGENINGEN UR
Animal Breeding & Genomics Centre

Acknowledgements

Chicken genome sequencing consortium
Chicken SNP consortium
Swine SNP consortium
Swine genome sequencing consortium

Funding
USDA
EU-Sabre
EU-Eadgene
Institute for pig genetics (IPG)
Hendrix Genetics

Thank You

ANIMAL SCIENCES GROUP WAGENINGEN UR
Animal Breeding & Genomics Centre

Acknowledgements (1)

Funding
- USDA – CSREES
- EU (Sabre, PigSNP)
- Institute for Pig Genetics (IPG)

Food Quality and Safety

DNA samples

- China Agricultural University
  - Ning Li
- University of Sassari, Italy
  - Massimo Scandura
- Staff Institute, Japan
  - Naohiko Okumura
- INRA-Toulouse, France
  - Alain Duvro
- MARC, USA
  - Gerry Rohrer
- Roslin Institute, UK
  - Alan Archibald
- EU PigBiodiv1
- PTP, Italy
  - Elisabetta Guiffra
- REPROGEN, Australia
  - Jaime Gongora
- Universidade de Lisboa, Portugal
  - Deodália Dias
- Universitat Autonoma Barcelona
  - Miguel Perez-Enciso
- University of Aarhus, Denmark
  - Christian Bendixen
- Durham University, UK
  - Greger Larsen
- Iowa State University, USA
  - Max Rothschild
- Wageningen University, The Netherlands
  - Richard Crooijmans
- University of Illinois, USA
  - Larry Schook
- Aristotle University of Thessaloniki, Greece
  - Costas Triantaphyllidis
- Istituto Zootecnico per la sardegna, Italy
  - Sara Casu
- Research Centre for Biology – LIPI, Indonesia
  - Gono Semiadi
- Hendrix Genetics (HG)
- Institute for pg genetics (IPG)
- Pig Improvement Company (PIC)

ANIMAL SCIENCES GROUP WAGENINGEN UR
Animal Breeding & Genomics Centre

Acknowledgements (2)

- Wageningen University, The Netherlands
  - Richard Crooijmans
  - Marcos Ramos
  - Andreia Fonseca
  - Hinri Kerstens
  - Hendrik-Jan Megens
- University of Aarhus, Denmark
  - Christian Bendixen
  - Jakob Hedegaard
- USDA,ARS, MARC, USA
  - Gary Rohrer
  - Tim Smith
  - Dan Nonneman
- INRA , Toulouse, France
  - Denis Milan
- Roslin Institute, UK
  - Alan Archibald
  - Andy Law
- Durham University, UK
  - Greger Larsen
- Purdue University, USA
  - Bill Muir
- University of Illinois, USA
  - Larry Schook
  - Jon Beever
- USDA, ARS, Beltsville, USA
  - Curt Van Tassel
- Iowa State University, USA
  - Max Rothschild
  - Zhiliang Hu
  - James Reecy
- Sanger Institute, UK
  - Carol Churcher
  - Richard Clark
- University of Missour-Columbia, USA
  - Jeremy Taylor
  - Bob Schnabel
- Universitat Autonoma Barcelona
  - Miguel Perez-Enciso
  - Lucca Ferreti
- Illumina Inc., San Diego, USA
  - Mark Hansen
  - Marylinn Munson

ANIMAL SCIENCES GROUP WAGENINGEN UR
Animal Breeding & Genomics Centre