UPPSALA
UNIVERSITET

Whole genome resequencing reveals loci under selection during chicken domestication

Carl-Johan Rubin

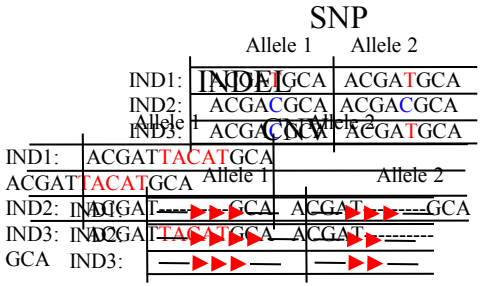
Dept. Med. Biochemistry & Microbiology
Uppsala University,
Sweden

UPPSALA UNIVERSITET

Genetic variation

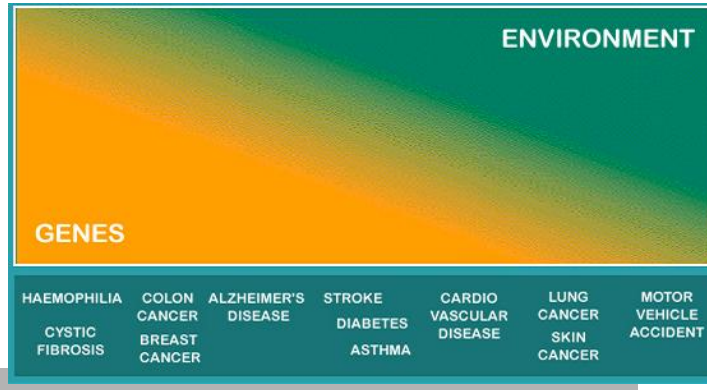
Different types of variation

- Human genome = 3 x 10^9 nucleotides, chicken genome = 1 x 10^9 nucleotides
  - Inter-individual variation exists in several forms:
    - 1) Single Nucleotide Polymorphisms (**SNPs**)
    - 2) One or more nucleotides **in**serted/**del**eted (small **indels**)
    - 3) **C**opy **N**umber **V**ariants (large stretches of DNA)



- In both chicken and human, such variable genomic elements explain phenotypic variation (in combination with environmental factors)

Important to identify the variants that confer differences between individuals / disease:
- Drug development, personalized medicine
- Screening for disease at early stage
- Maximize yields from animals/plants



UPPSALA UNIVERSITET

Domestic Animal Genetics

From domestication to the dissection of complex traits

Domestic animals have often been artificially selected for certain traits over several thousand years










Excellent models for deciphering the genetics of complex traits

UPPSALA UNIVERSITET

Domestic Animal Genetics

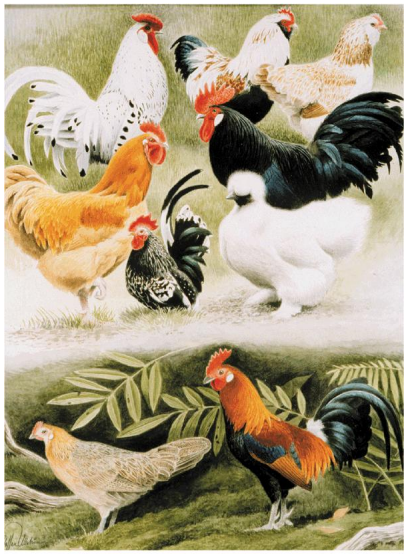
From domestication to the dissection of complex traits

- Domestication - selection based to result in man's benefit (Darwin, 1859)
  - Relaxation of natural selection pressures (predation/ starvation)
  - Intensified selection of traits preferred by humans
  - Natural selection under captivity → adaptation
- Domestication leads to "Domestic Phenotype"
  - External morphology - colour, fur, body size, smaller skulls and legs (Clutton-Brock, 1998)
  - Internal morphology -↓ in brain size, smaller intestines (Kruska, 1996)
  - Physiological changes - endochrine response, reproductive cycle (Setchell, 1992)
  - Developmental changes - earlier sexual maturity (Belyaev, 1984)
  - Behaivoural changes - ↓ fear, ↑sociability, ↓antipredator response (Price, 1997)
- To elucidate modern phenotypes, look to domestication where millennia of selection has created the perfect models

Chicken domestication

UPPSALA UNIVERSITET

- Process began ~8,000 years ago in South Asia
  - Mainly Red junglefowl, some contribution from the Grey junglefowl
- Domestic varieties phenotypically more diverse
  - Plumage colour, production purpose and traits



Nature Reviews | Genetics

Chicken domestication

UPPSALA UNIVERSITET

- Process began ~8,000 years ago in South Asia
  - Mainly Red junglefowl, some contribution from the Grey junglefowl
- Domestic varieties phenotypically more diverse
  - Plumage colour, production purpose and traits

Nature Reviews | Genetics

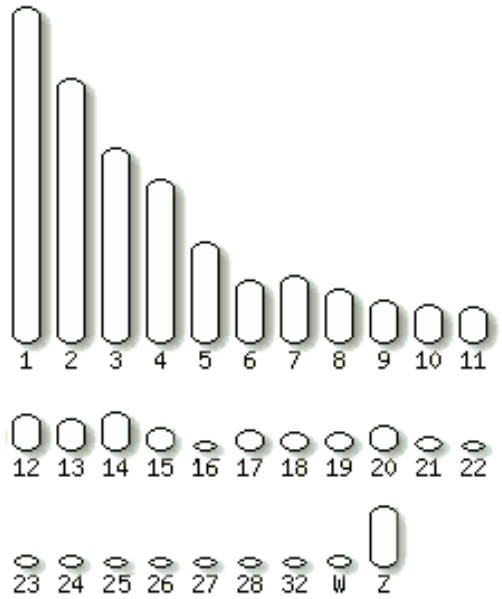
- For a long time domestic chicken were multi-purpose
- Last century, intensive selection → **Many** specialized domestic breeds
- Most important source of animal protein worldwide

Previous genome studies in chicken

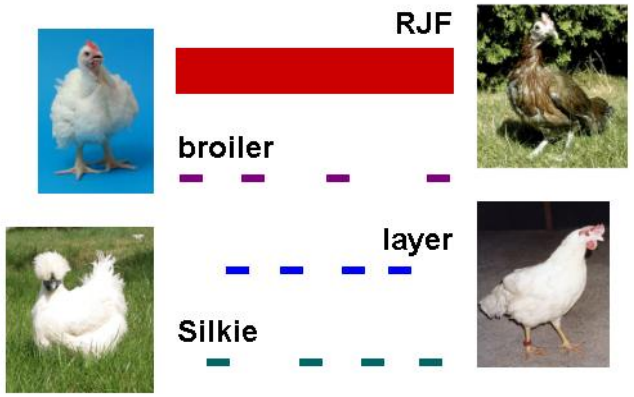
Draft assembly of chicken genome:

6.6 x coverage of red junglefowl genome (1.1 Gb)

Hillier et al. 2004 Nature, 432:695-716.



A genetic variation map for chicken with 2.8 million SNPs
International Chicken Polymorphism Map Consortium
Wang et al. 2004 Nature 432:717-722



Compare partial sequences with red junglefowl genome

Broiler = **0.25X** coverage

Layer (White Leghorn) = **0.25X** coverage

Silkie = **0.25X** coverage

Identified SNPs have been valuable in SNP typing studies, but no domestication loci identified. Resolution?

UPPSALA UNIVERSITET

DNA-sequencing revolution

- Over last 4-5 years technical advances have revolutionized DNA sequencing

- Novel techniques enable sequencing of 50-100 x 10^9 DNA bases / machine week

- Cost per base decrease: 10x each year
- Sequence output increase: 10x each year

- "Old way" (Sanger sequencing) → Few long reads
- "New technologies" → Many short reads

- Chicken genome is small (1/3 of mammals) and has few repetitive motifs --> perfect species for genome resequencing in 2008.

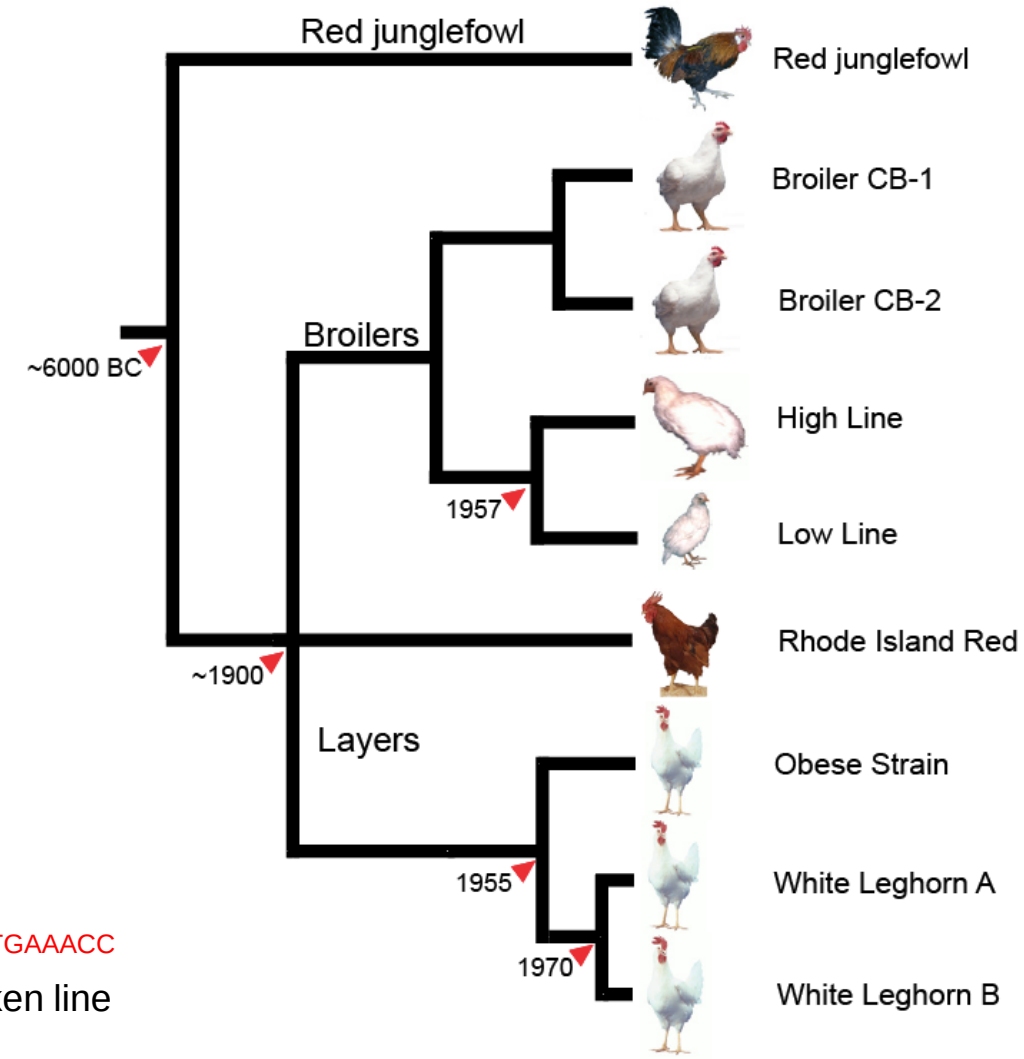
UPPSALA UNIVERSITET

Our approach – whole genome resequencing

- Scope of experiment
  - Sequence **pools** of chicken lines (domestic & wild)
  - Identify most common allele at all SNP positions
  - Identify fixed genetic differences between divergent groups
  - **Use AB SOLiD chemistry:**
    - 35 bp reads,

      AGACTCGTACCGAGAGATAGTCTCTCCATGAAACC

      4-5 x genome coverage / chicken line



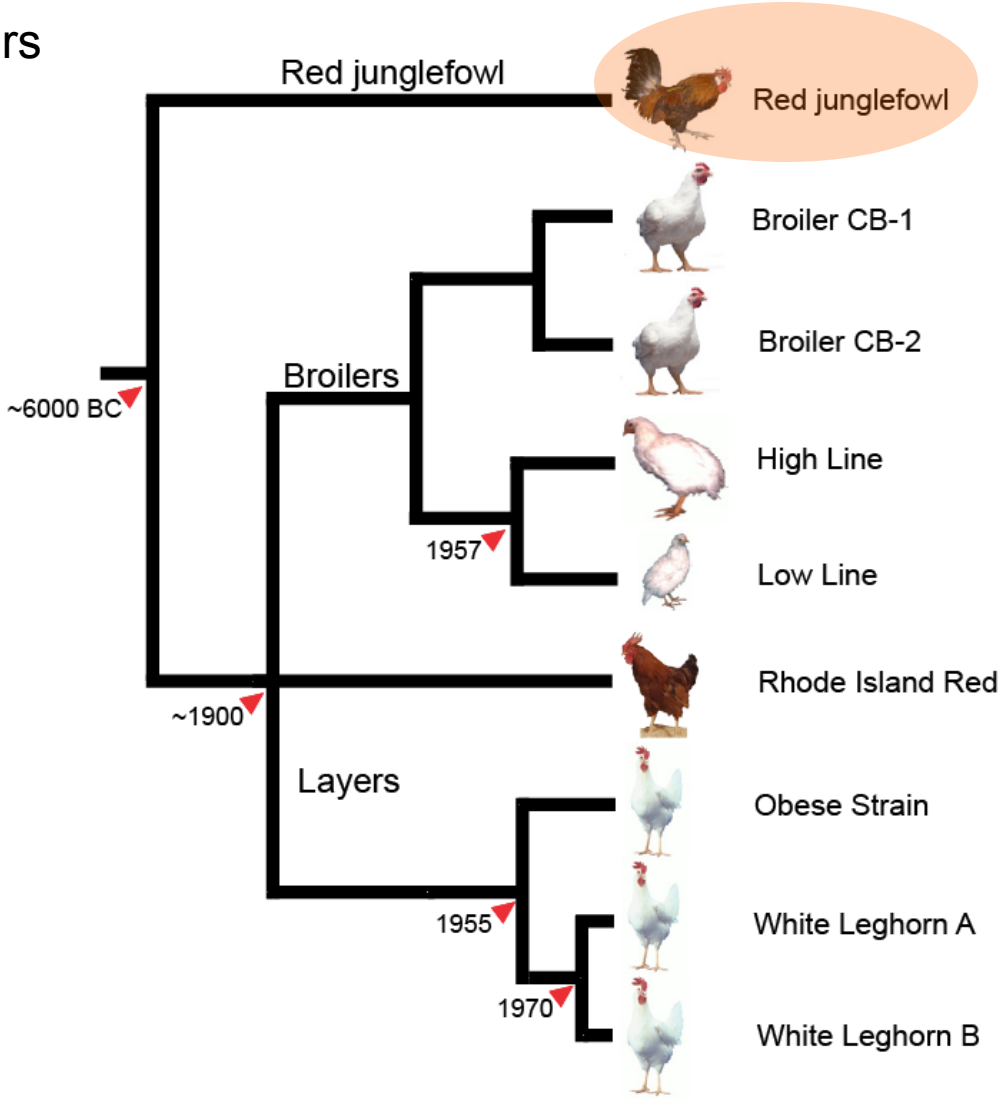
Diverse DNA pools

"Wild", Layers and Broilers

Red junglefowl

RJF-pool
2 zoo populations
8 males
Sweden

RJF-ref
Reference bird
Female
Partially inbred UCD 001



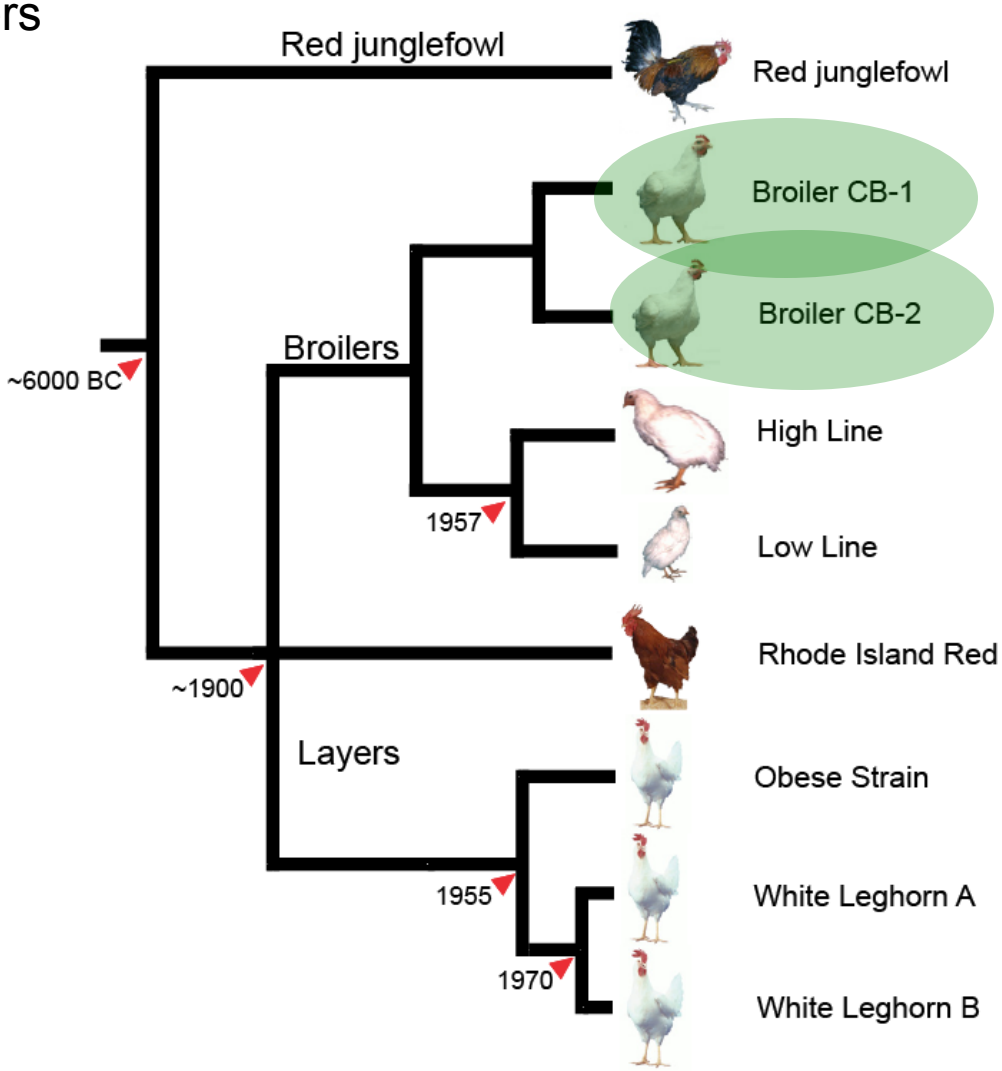
Diverse DNA pools

"Wild", Layers and Broilers

Broilers

CB-1
Commercial broiler line
10 males
Sweden

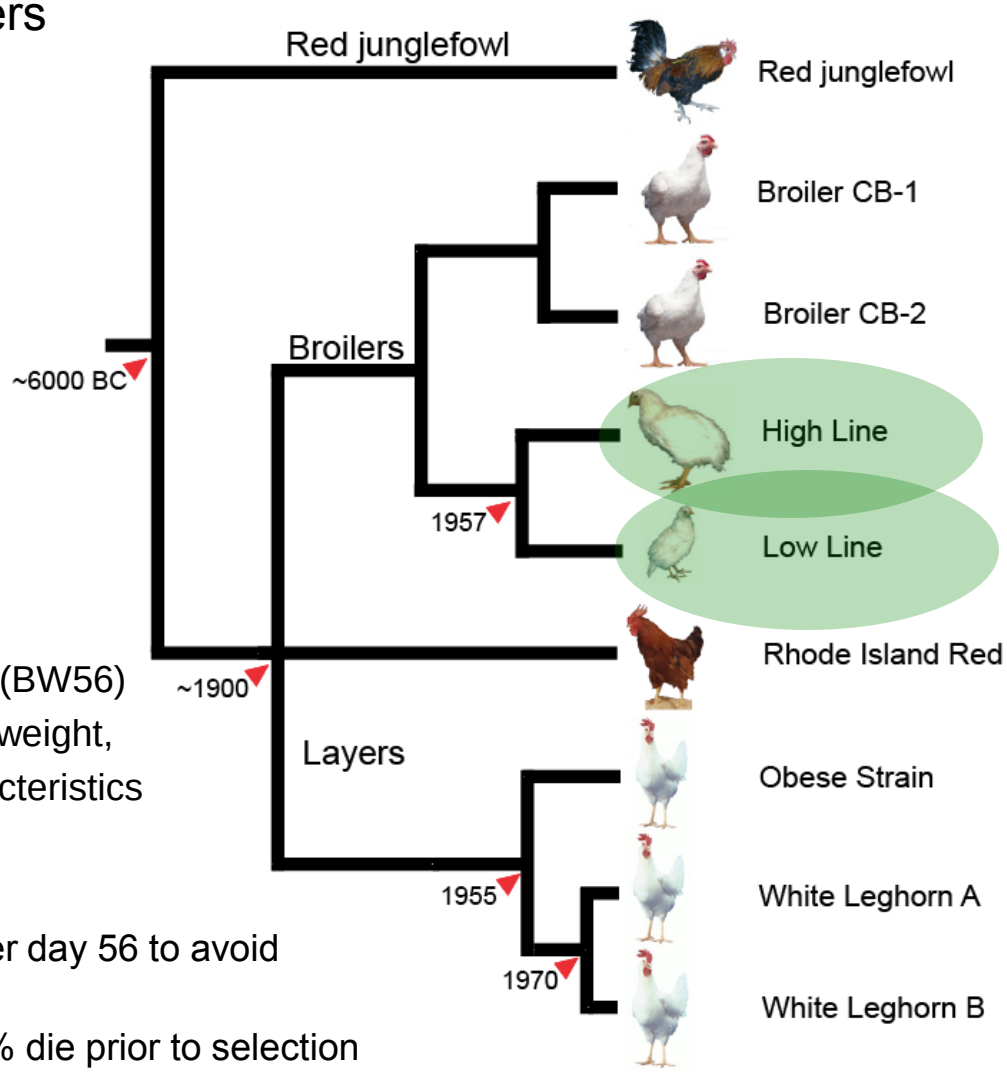
CB-2
Commercial broiler line
10 females
France



UPPSALA UNIVERSITET

Diverse DNA pools

"Wild", Layers and Broilers



Broilers

High Growth Line

7 males, 4 females

USA

Low Growth Line

7 males, 4 females

USA

Divergently selected on body weight at 56 days (BW56)

After 40 generations, 9x difference in body weight, growth rate, fat deposition, metabolic characteristics

Appetite :

"High" increased, must be fed restricted diet after day 56 to avoid metabolic disease

"Low" reduced, display anorexic phenotype, 20% die prior to selection

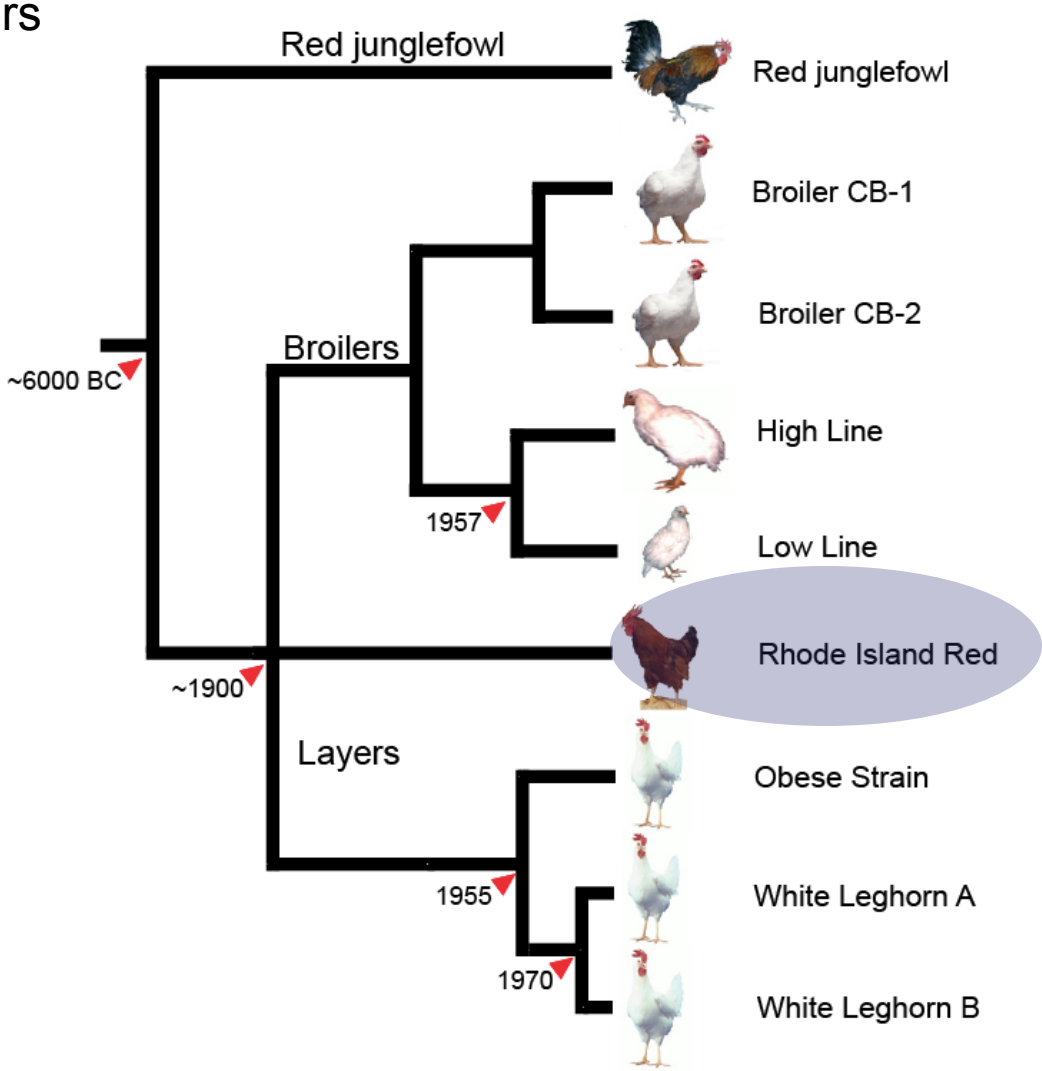
UPPSALA UNIVERSITET

Diverse DNA pools

"Wild", Layers and Broilers

Layers

- Rhode Island Red
- Commercial population
- 8 males
- France



UPPSALA UNIVERSITET

Diverse DNA pools

"Wild", Layers and Broilers

Layers

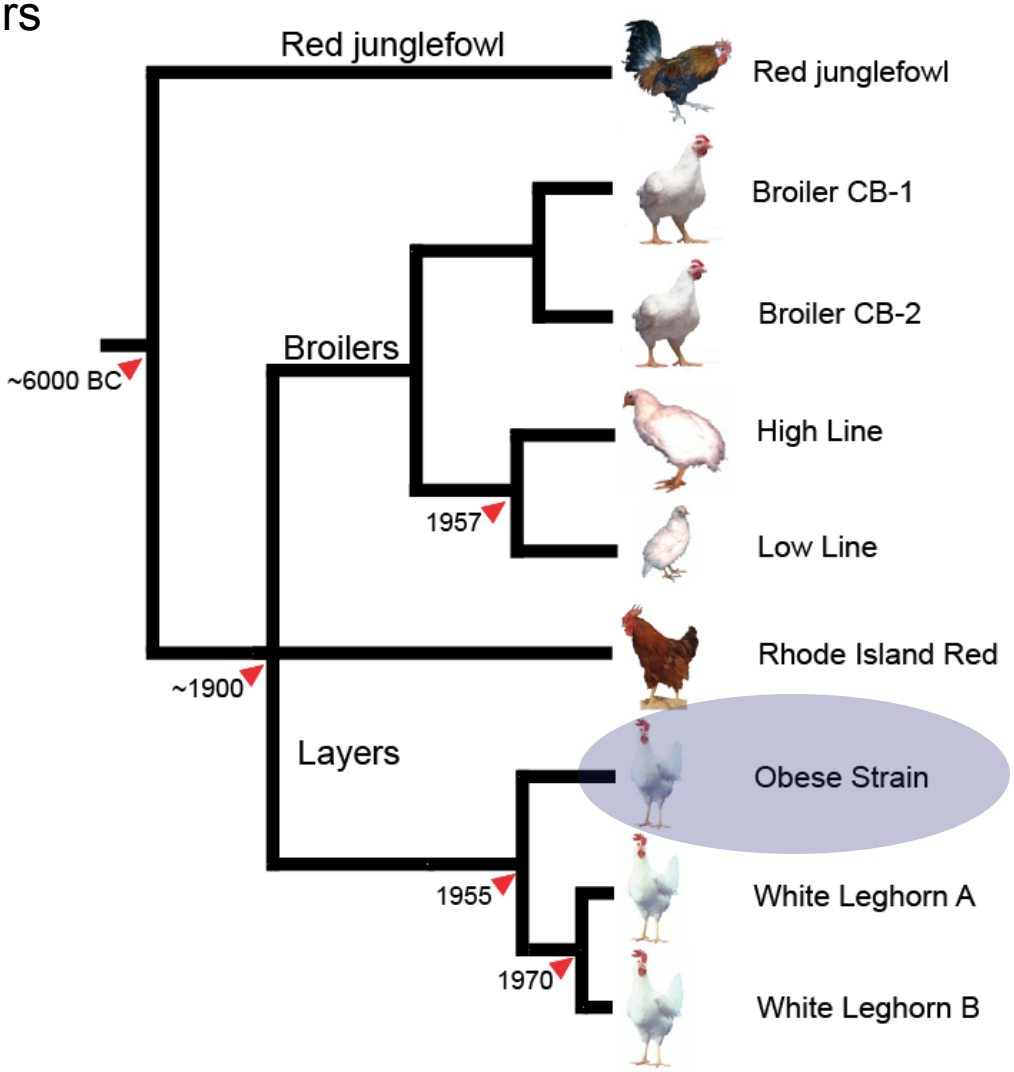
Obese Strain

Developed 1955 as model for autoimmune thyroiditis

Taken from White Leghorn line

10 males

USA



Diverse DNA pools

"Wild", Layers and Broilers

Layers

White Leghorn A

White Leghorn Line 13

11 males

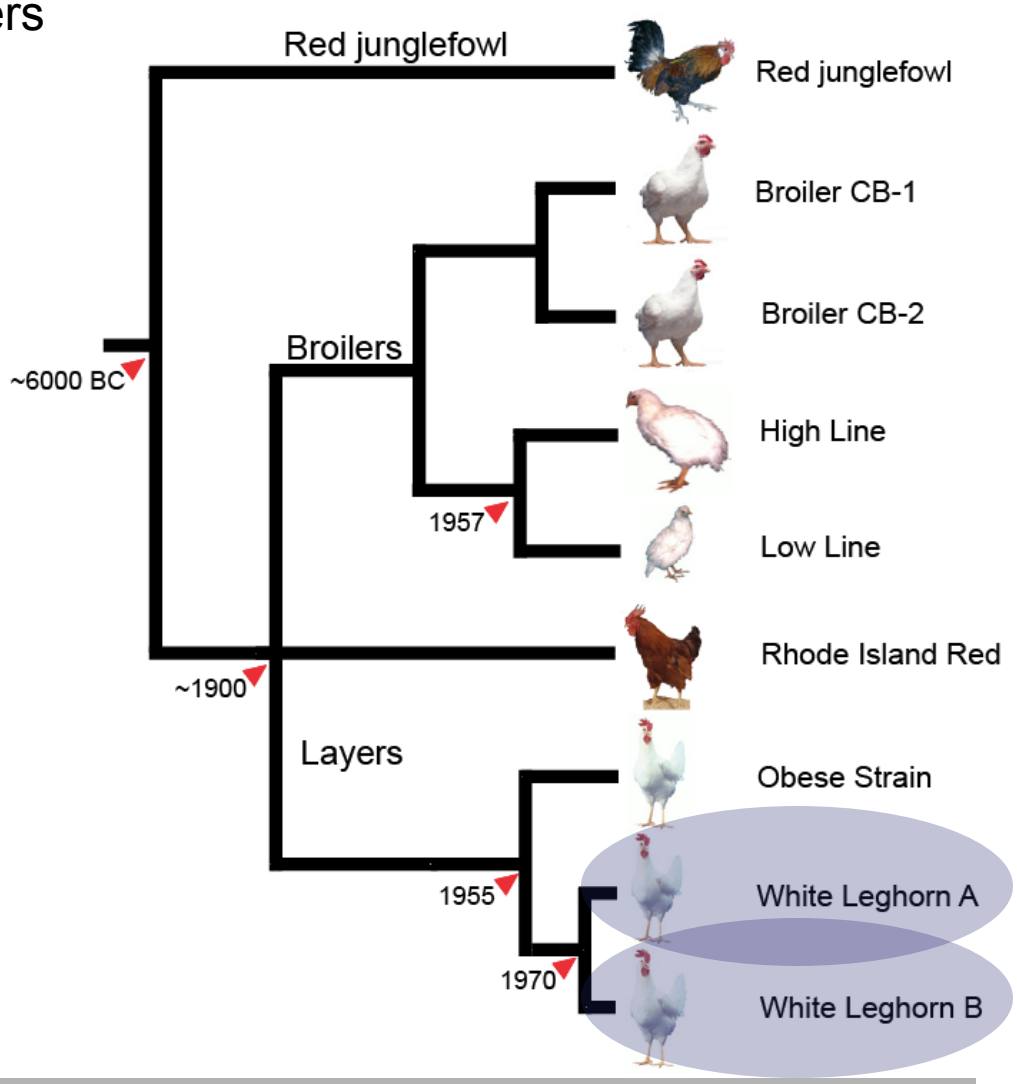
Sweden (SLU)

White Leghorn B

Commercial White Leghorn

8 males

USA

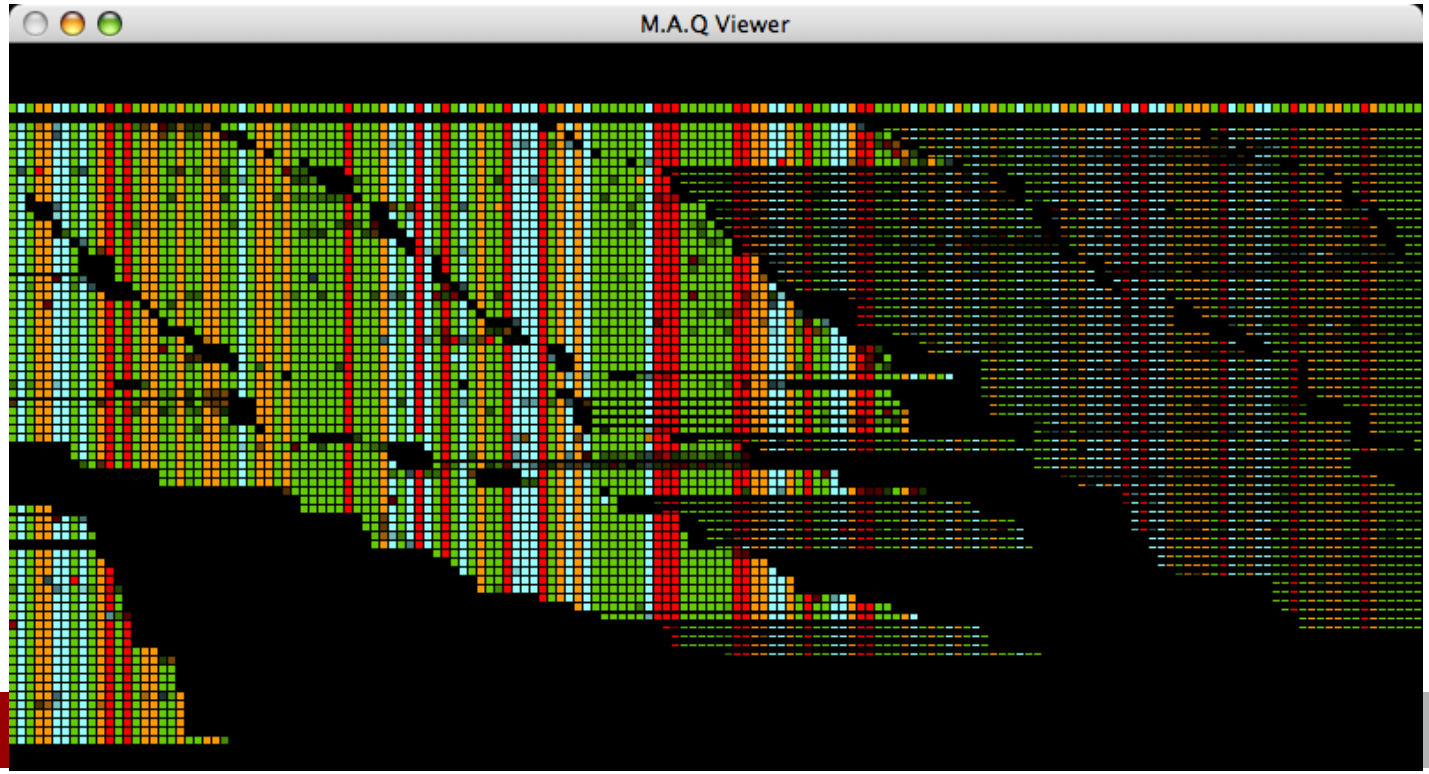


UPPSALA UNIVERSITET

Step 1: Alignment of reads

- Sequence reads were aligned (mapped) to reference genome framework (the red junglefowl complete genome sequence)

- Mapping performed with Corona Lite software (Life Technologies), *criteria <= 3 differences tolerated*



UPPSALA UNIVERSITET

Step 1: Alignment of reads

- Sequence reads are aligned (mapped) to reference genome framework (published red junglefowl genome sequence)
- Mapping performed with Corona Lite software (Life Technologies) *criteria <= 3 differences tolerated*
- 10^8 – 10^9 sequence reads (35 bp) searched against 10^9 nucleotides in reference genome (computationally intense step)
  - Only best unique matches will be retained

    If several placements equally good, can´t tell which is correct
  - Algorithm must account for potential sequence variation (SNPs) or sequence errors in read

UPPSALA UNIVERSITET

Alignment statistics

	Line	Gb Aligned	Mb Covered	Coverage
LAYERS	White Leghorn A (WL-A)	2.75	818	3.37
	White Leghorn B (WLH-B)	3.41	852	4.00
	Obese Leghorn (OS)	2.99	828	3.61
	Rhode Is Red (RIR)	4.58	885	5.18
BROILERS	Commercial Broiler 1(CB1)	3.35	835	4.01
	Commercial Broiler 2 (CB2)	2.65	800	3.32
	High Growth Line (High)	4.57	882	5.19
	Low Growth Line (Low)	4.90	887	5.53
WILD	Red Junglefowl (RJF-Pool)	6.28	904	6.95
	Red Junglefowl (RJF-Ref)	2.70	809	3.34

UPPSALA UNIVERSITET

Alignment statistics

	Line	Gb Aligned	Mb Covered	Coverage
LAYERS	White Leghorn A (WL-A)	2.75	818	3.37
	White Leghorn B (WLH-B)	3.41	852	4.00
	Obese Leghorn (OS)	2.99	828	3.61
	Rhode Is Red (RIR)	4.58	885	5.18
BROILERS	Commercial Broiler 1(CB1)	3.35	835	4.01
	Commercial Broiler 2 (CB2)	2.65	800	3.32
	High Growth Line (High)	4.57	882	5.19
	Low Growth Line (Low)	4.90	887	5.53
WILD	Red Junglefowl (RJF-Pool)	6.28	904	6.95
	Red Junglefowl (RJF-Ref)	2.70	809	3.34

- 80% of genome covered – 145 Mb missing repetitive seq, artificial duplications, or underrepresented at lab. amplification step

UPPSALA UNIVERSITET

Step 2: SNP identification

Identifying SNPs from sequence data

- **SNP** calling performed with software Corona Lite (Life Technologies)
  - >= 3 independent reads showing same non-reference nucleotide → SNP
  - (Indels and CNVs cannot be identified when 35bp reads are analyzed)

- Total of 7,493,903 unique variant loci from combined data
  - One every 133 bp in the genome

UPPSALA UNIVERSITET

Step 2: SNP identification

Identifying SNPs from sequence data

- SNP calling performed with software Corona Lite (Life Technologies)
  - >= 3 independent reads showing same non-reference nucleotide → SNP
  - (Indels and CNVs cannot be identified when 35bp reads are analyzed)

- Total of 7,493,903 unique variant loci from combined data
  - One every 133 bp in the genome

- **Low false positive error rate**
  - 3 – 5 million non-reference alleles found in individual pooled populations
  - Only 218, 662 non-reference alleles found in reference genome bird

    Large proportion of these regions, ref bird that is heterozygous (1 allele in genome assembly)

  - On Z chromosome there should be no variation in reference bird.

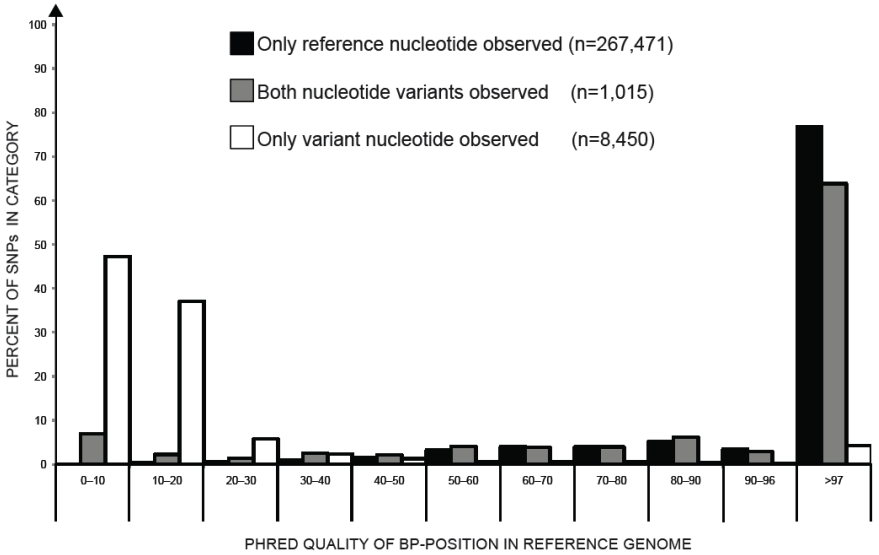
    Data reference bird: reference base = 96.6%, non-ref = 3.0%, both bases = 0.4%

UPPSALA UNIVERSITET

> 7 million unique SNPs found

One every 133 bp in the genome

- SNP filtering to eliminate putative SNP where
  - Only non-ref allele called
  - > 1 read from ref bird
  - Quality of genome ref sequence < 50



- Pruned 40,058 putative SNP

- 7,453,845 confident SNP taken forward to analysis.

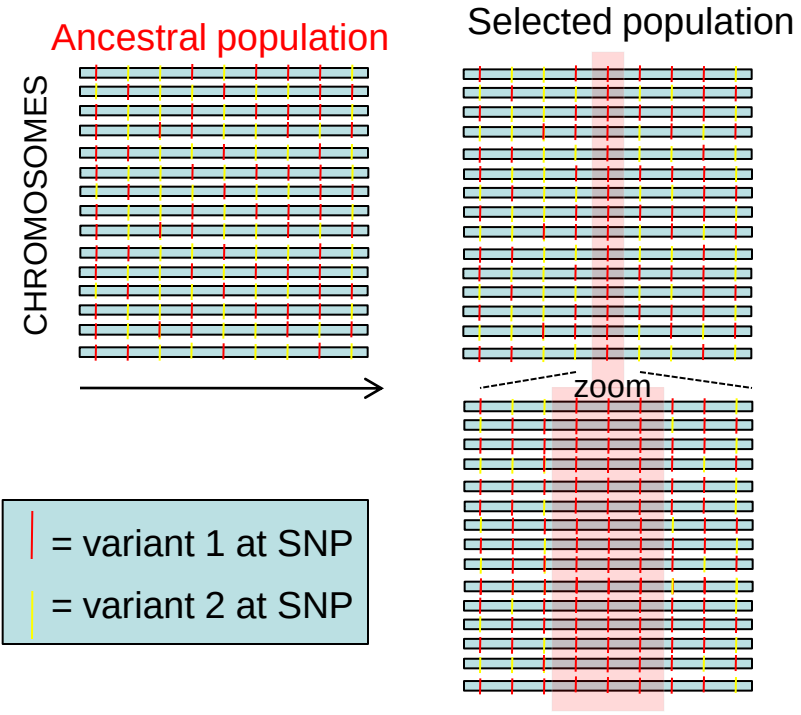
- Validation study of 321 SNPs suggests 98.8 % of these are correct SNPs

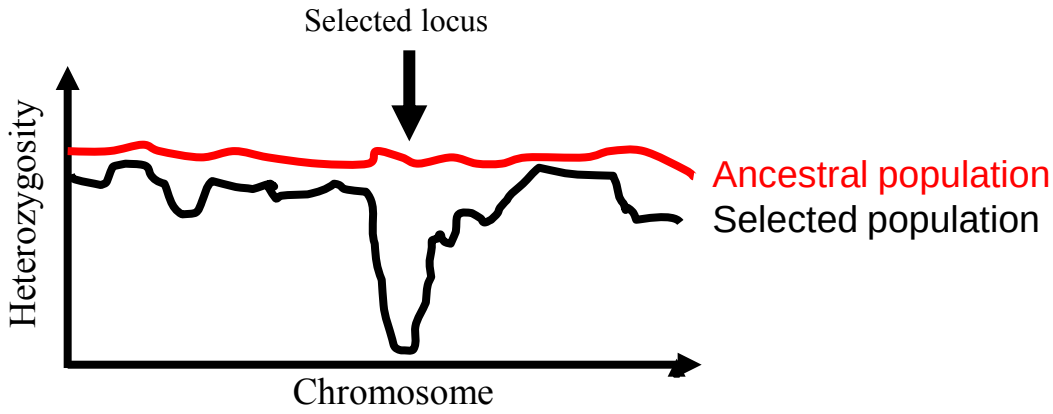
UPPSALA UNIVERSITET

Selective Sweeps

Searching for highly fixed genomic regions

– **Selective sweep:**
  – Region of genome that has reduced or no variation due to selection of a beneficial variant
  – Beneficial variant will carry with it other "fixed variants" → sweep



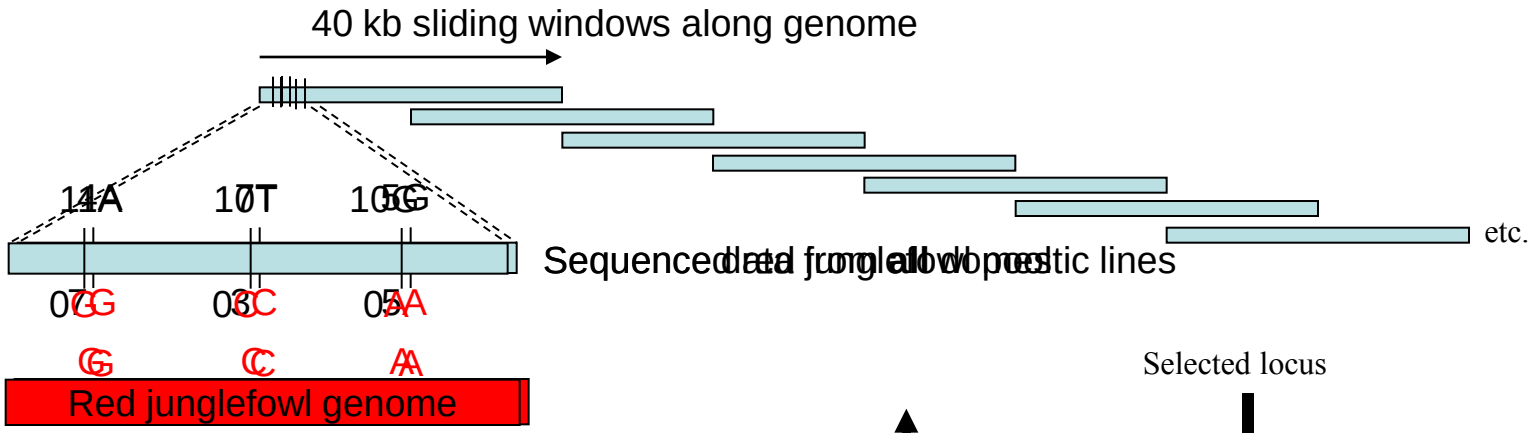


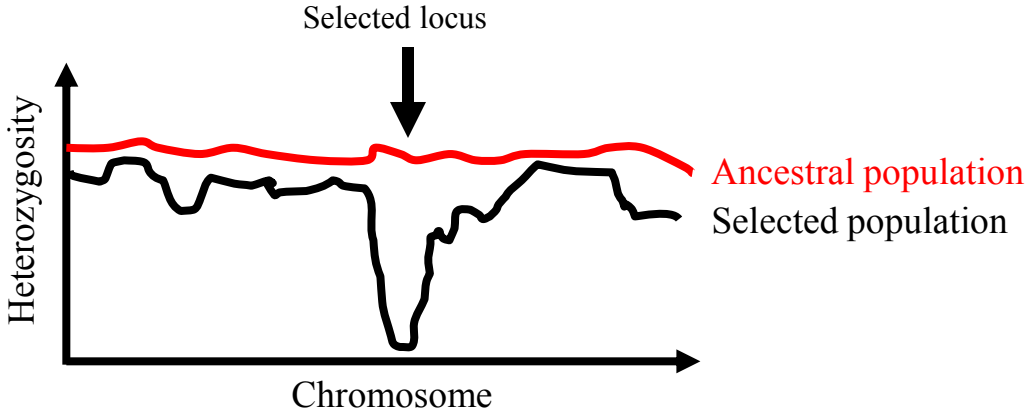
– **Method:** Analyze all SNP data (7.5 million) from breeds selected for similar traits
  → selected regions should be shared (identical by descent)

Selective Sweeps

Searching for highly fixed genomic regions

Allele counts for all reads overlapping SNPs are assessed





UPPSALA UNIVERSITET

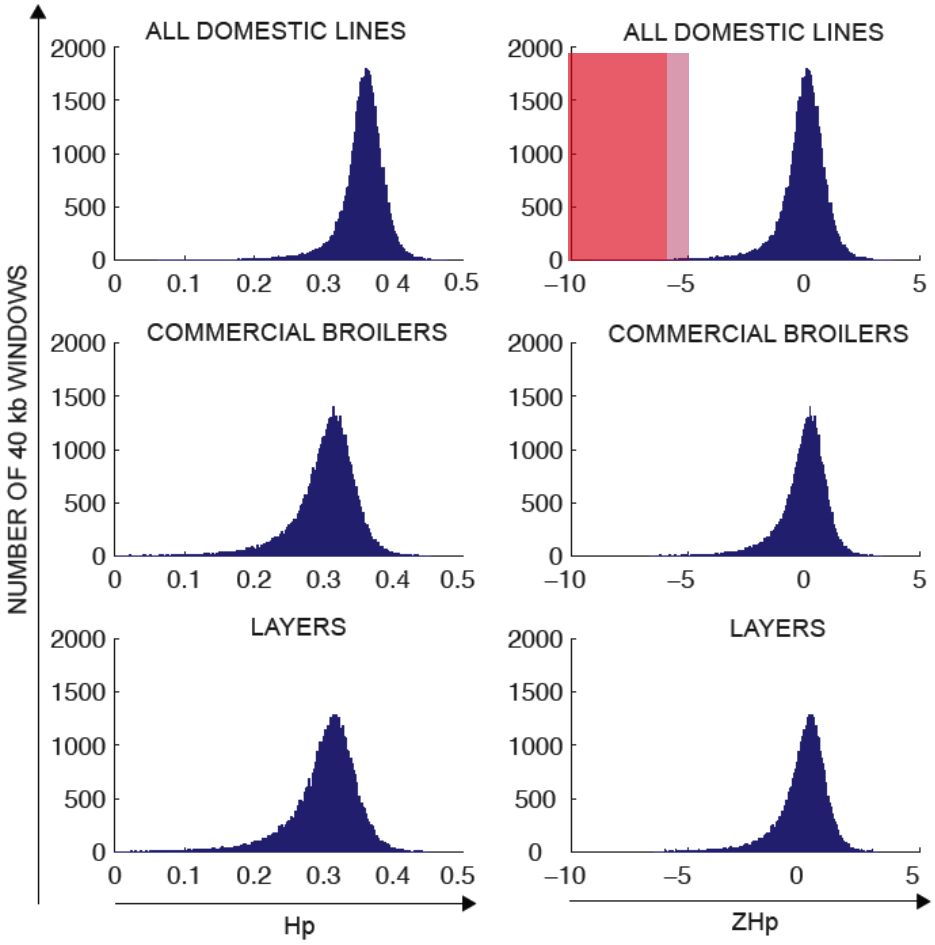
Selective Sweeps

Searching for high fixation: selective sweep or genetic drift?

- Chr 1-28 = 47,808 windows
- Assume ZH_p normal distribution

  Values < -5 significant

  Our cut off ZH_p < -6



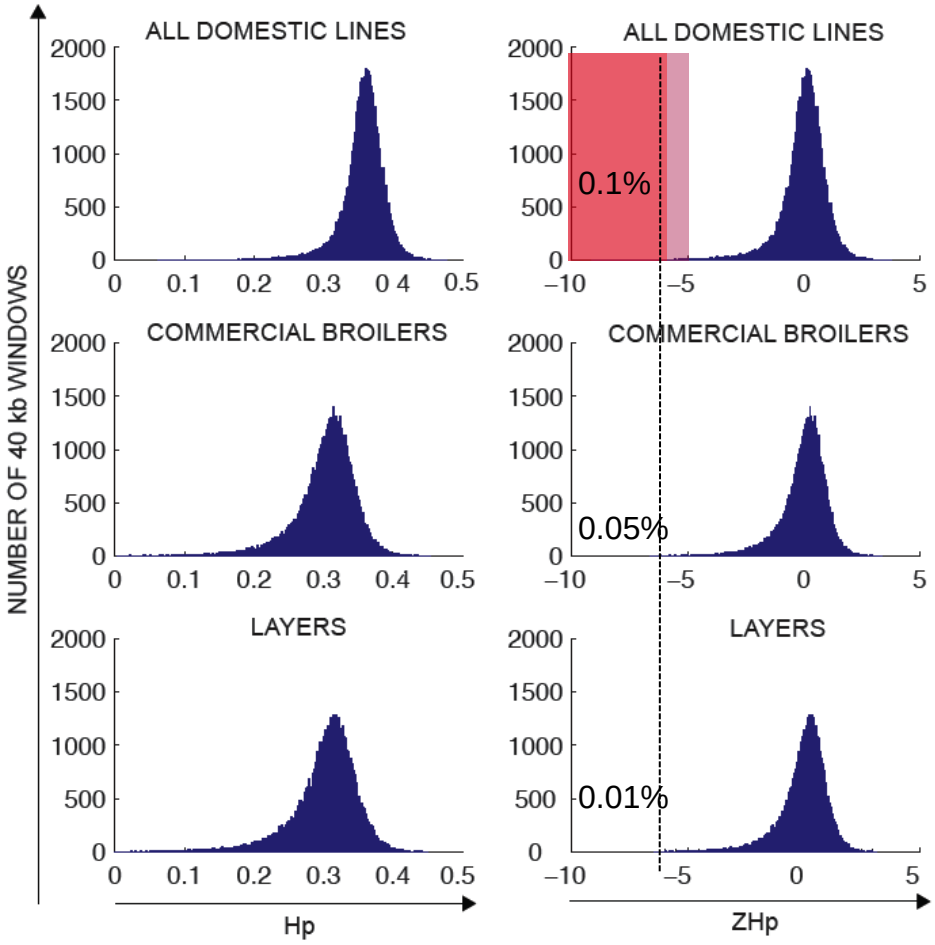
Selective Sweeps

Searching for high fixation: selective sweep or genetic drift?

- Chr 1-28 = 47,808 windows
- Assume ZH_p normal distribution

  Values < -5 significant

  Our cut off ZH_p < -6



Hypothesis:

True selective sweeps overrepresented among low ZHp-regions

Red junglefowl

Selective Sweeps

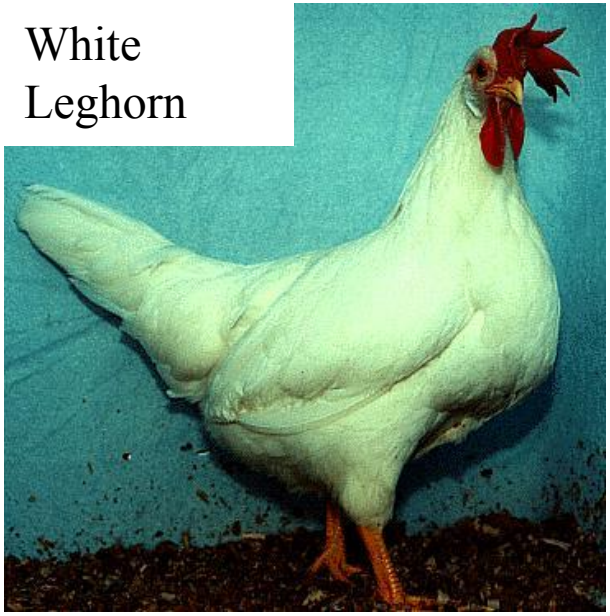
Known positive control selective sweep

Yellow skin predominates in domestic chicken

W: affecting metabolism of carotenoids

W+/- = white skin

White Leghorn



w/w = yellow skin

W first described by Bateson 1902

In 2008 J. Eriksson showed that yellow skin is caused by genetic variation in BCDO2 gene on chr. 24

UPPSALA UNIVERSITET

Selective Sweeps

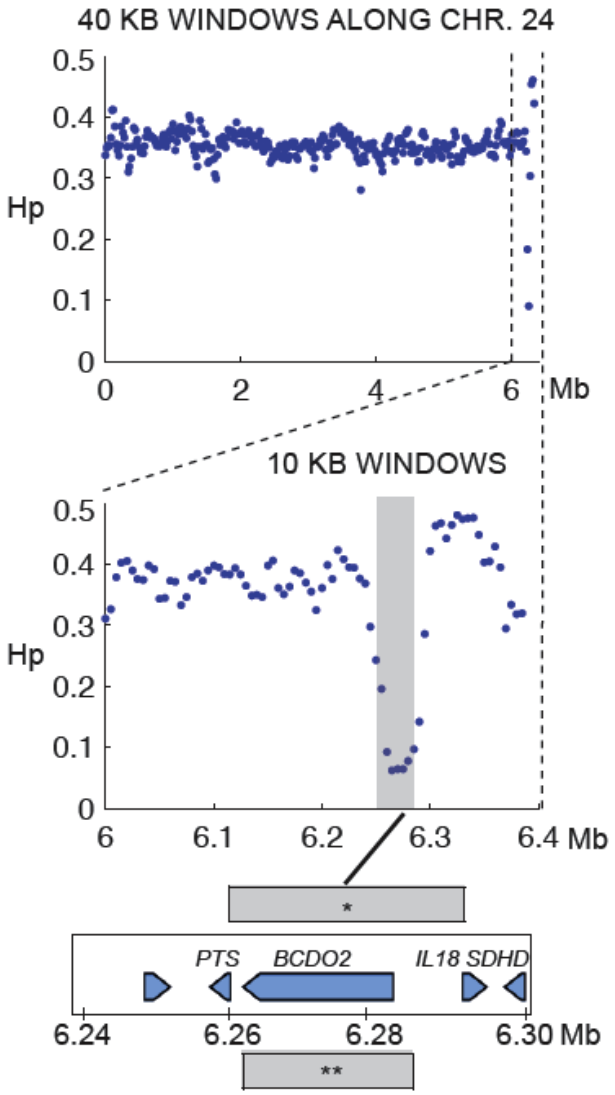
BCDO2: Proof of Principle

- *BCDO2* well established sweep
- Assume “All Dom” homozygous for *yellow skin*

- Sharp decline in Heterozygozity (H_p)
  - H_p = 0.09 over *BCDO2*
  - 40 kb sweep, overlaps with known region

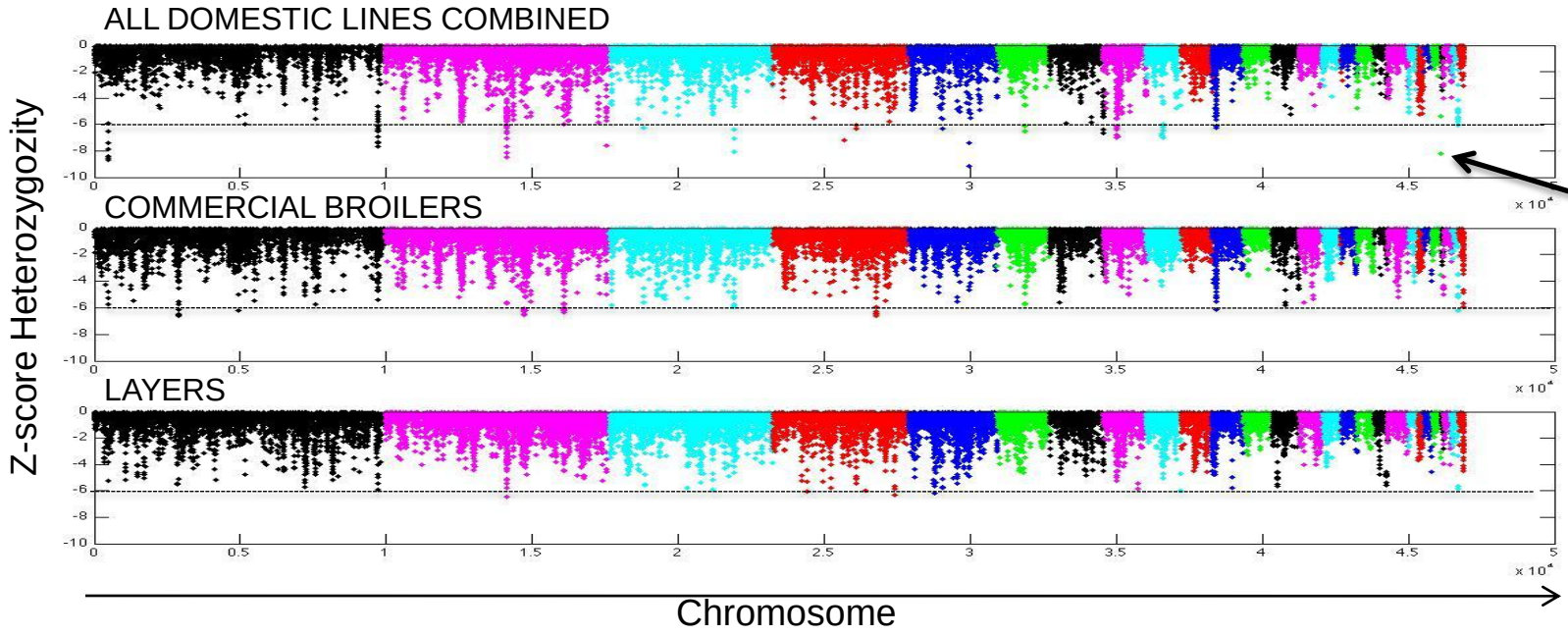
  - Not complete fixation,
    some birds carried wild type haplotype

- Demonstrates power to detect sweeps
- H_p high for large part of genome



Selective Sweeps

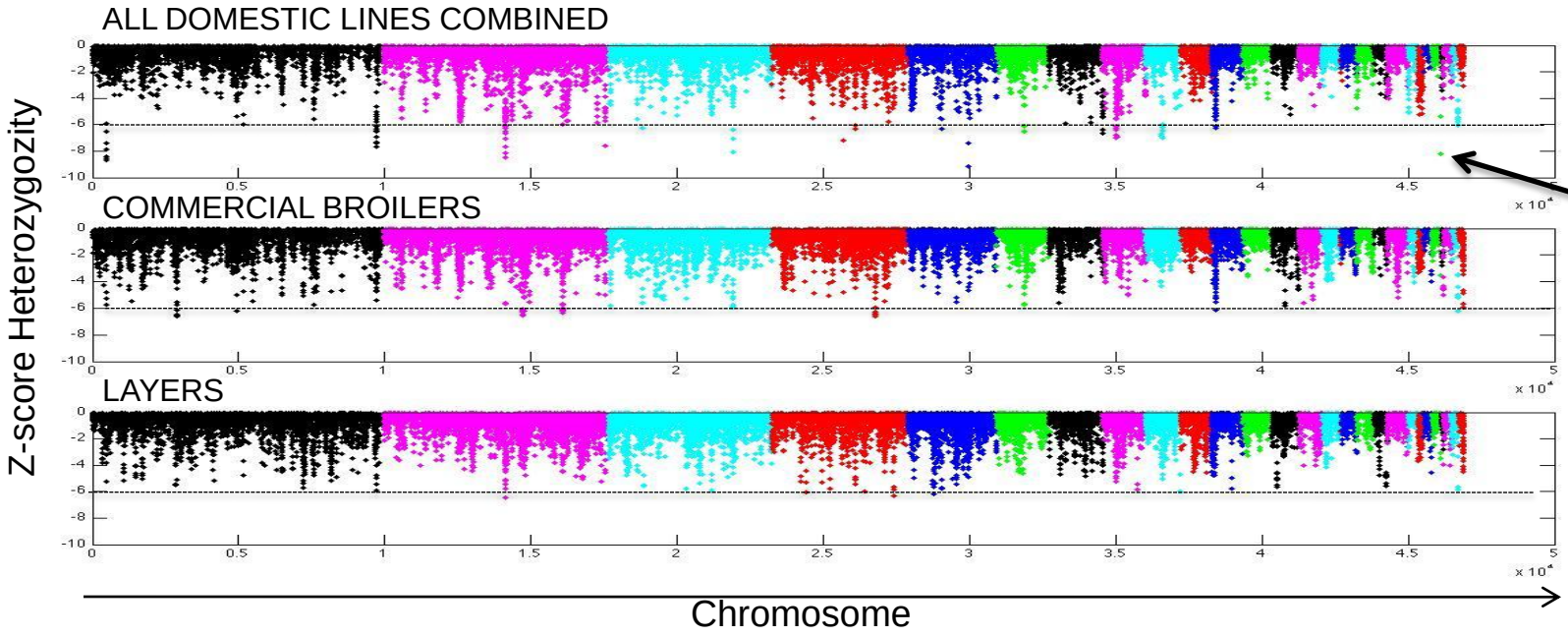
Global picture



UPPSALA UNIVERSITET

Selective Sweeps

Possible Sweeps in All Domestic



- 23 loci in “All Domestic” had one or more windows $ZH_p < -6$
- 3 loci as high or higher support than *BCDO2* ($ZH_p < -6$)
- Most candidate sweeps spans single genes

UPPSALA UNIVERSITET

Utility of Experimental Design

Take home messages

- These results important to chicken genomics
  - We identified > 7 million high quality SNP
  - We identified 40, 000 single base sequencing errors

- Chicken as a model for biomedical research
  - High throughput sequencing + domestic animal phenotypic selection
    = **high resolution identification of loci under selection**

- Chicken as production animal
  - 50 yrs of selection → chicken most important source of animal protein worldwide
  - We have identified some of the loci required to transform the red junglefowl into a highly efficient production animal
  - Results could be used to maintain biodiversity while maximizing production

UPPSALA UNIVERSITET

Design Not Limited to Chicken

- Price will continue to drop, sequence yield and length will rise
  - Apply same pooling technique to mammalian domestic species
  - Apply to natural populations where expect sweeps (e.g. environmental adaptation

- Current projects
  - More chicken sequencing:
    - Additional populations/lines
    - deeper coverage
    - longer reads 50 bp
    - paired end reads (2 x 50 bp reads separated by certain distance)

UPPSALA UNIVERSITET

Acknowledgements

UPPSALA UNIVERSITY

Michael Zody
Jonas Eriksson
Jennifer Meadows
Lin Jiang
Matt Webster
Max Ingman
Sojeong Ka
Finn Hallböök
Kerstin Lindblad-Toh
Leif Andersson

KAROLINSKA INSTITUTE

Ellen Sherwood

LINKÖPING UNIVERSITY

Per Jensen

SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES

Francois Besnier
Örjan Carlborg

BROAD INSTITUTE, USA

Ted Sharpe

INRA, FRANCE

Bertand Bed'hom
Michèle Tixier-Boichard

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY, USA

Paul Siegel