Developments in the detection of genetic variation (SNPs, CNVs and re-sequencing of genomes)

Richard Crooijmans

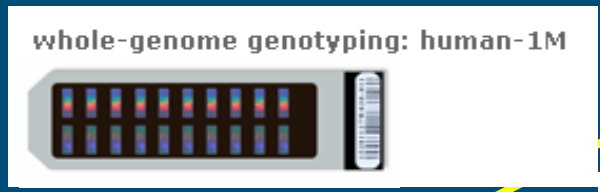
ANIMAL SCIENCES GROUP
WAGENINGEN UR

Animal Breeding & Genomics Centre

Overview

- SNP typing: Current status
  - Examples in chicken
- SNP discovery using 2nd generation sequencing technology (Solexa)
- Copy Number Variation (CNV)
- Re-sequencing of genomes

SNP genotyping assays



10^6 SNPs

100-500K SNPs

60K SNPs

1536 SNPs

1-384 SNPs

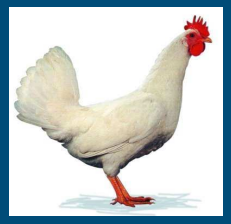
custom made

human

Chip design chicken 20k

- 3.5 million SNPs available (dbSNP)
- 2.8 million derived from comparison of sequence reads from three breeds against genome assembly (BGI-SNPs)


Broiler

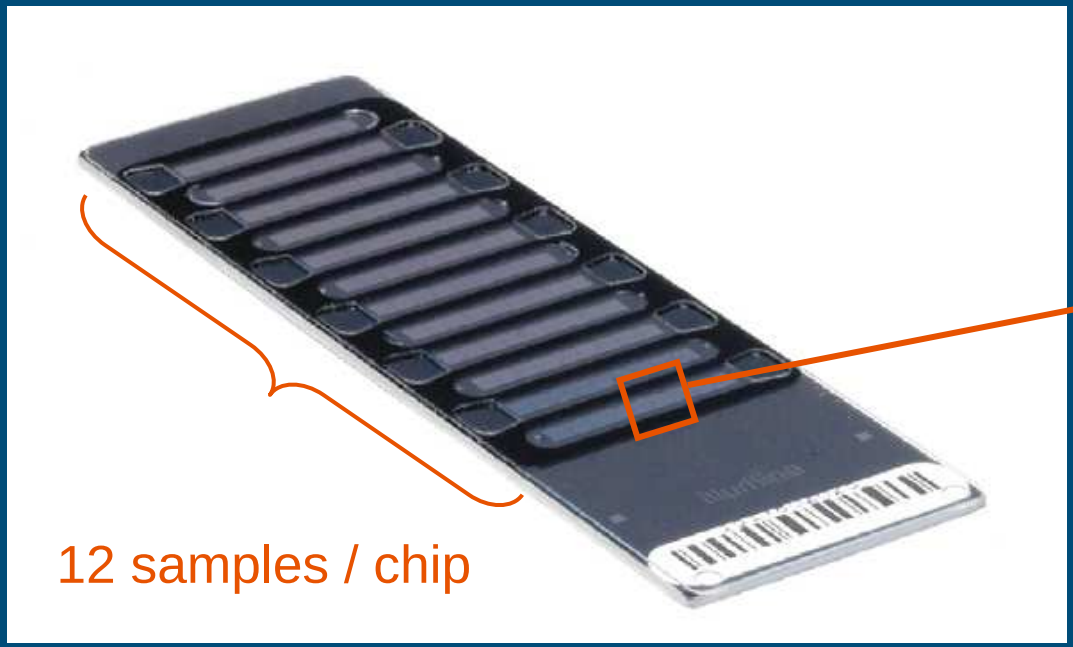

Layer

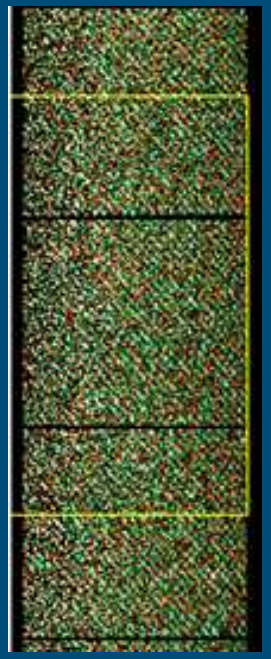

Silkie


RJF

- Information available for ~ 15,000 SNPs typed using the Illumina Golden Gate assay (Sentrix)

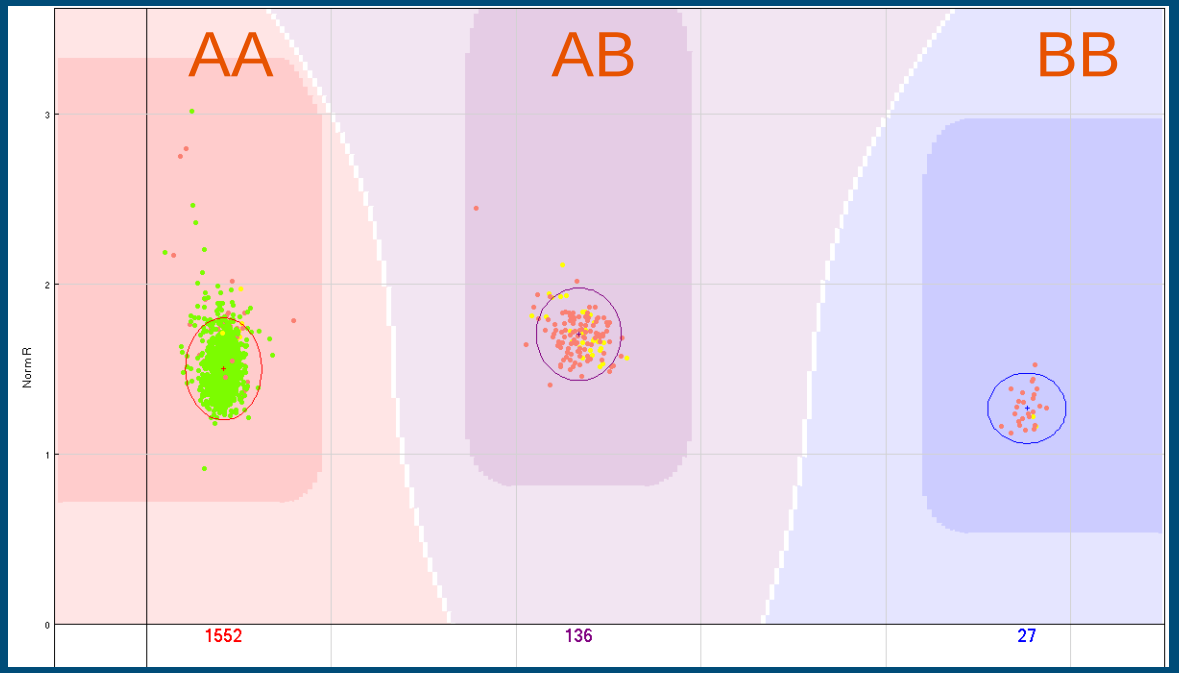
The infinium assay

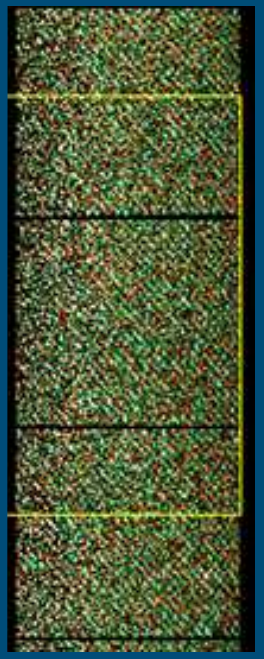




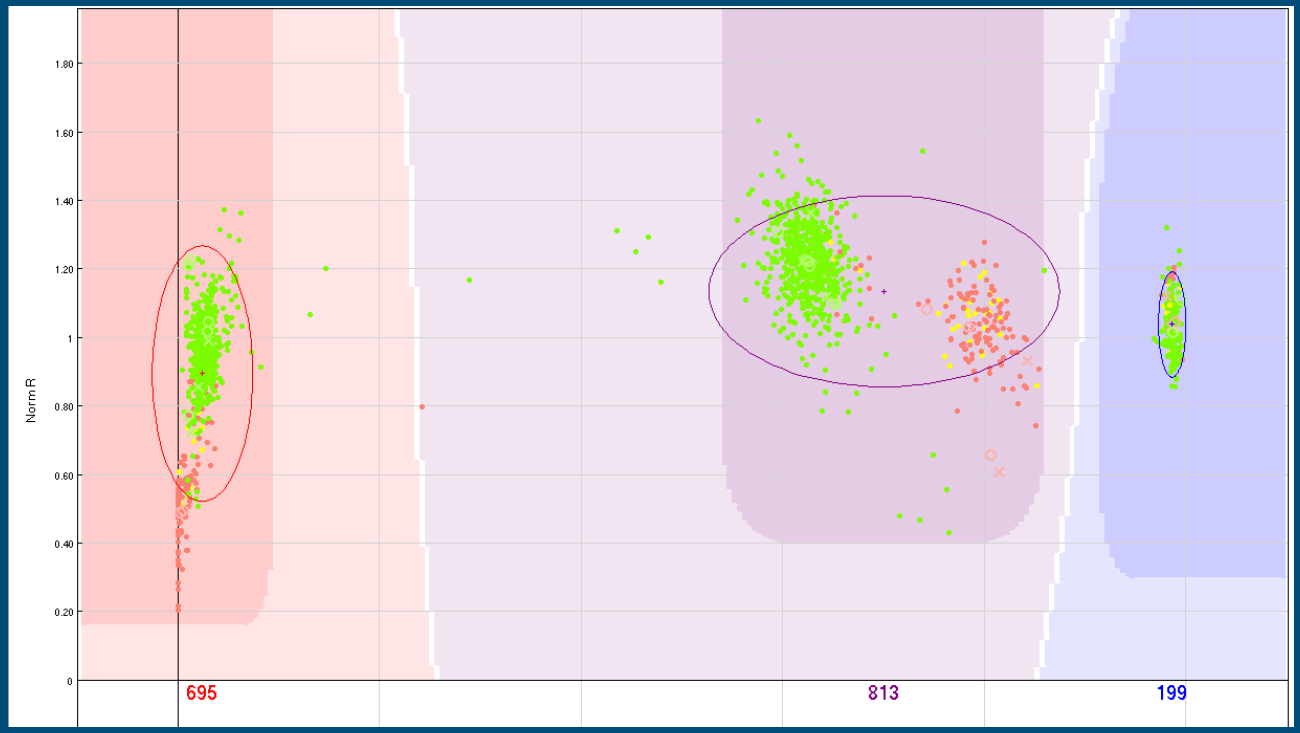
High call rates

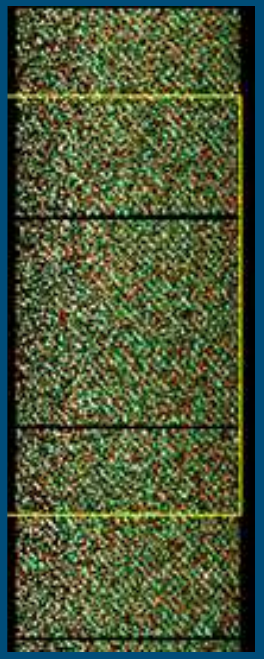
- Final chip contained 18,264 SNPs
- Genotytping calls for 17,790 SNPs (97.4 %)





Many SNPs with multiple heterozygous clusters



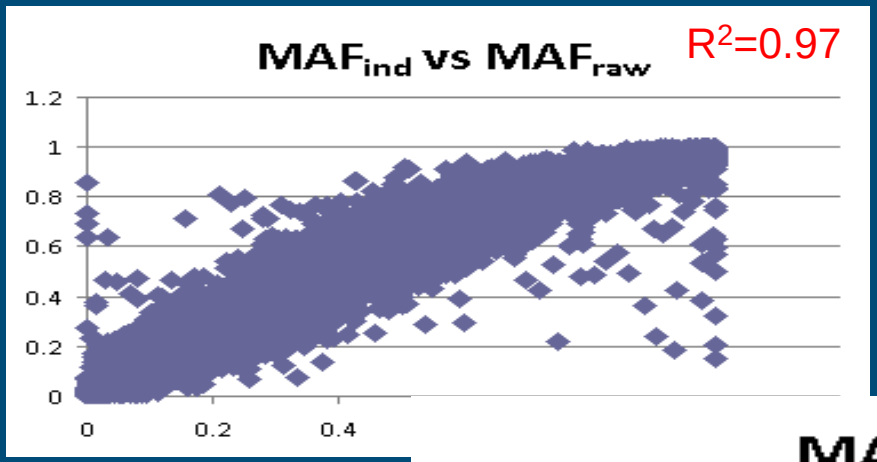


- Example of a SNP with population dependent clustering of alleles

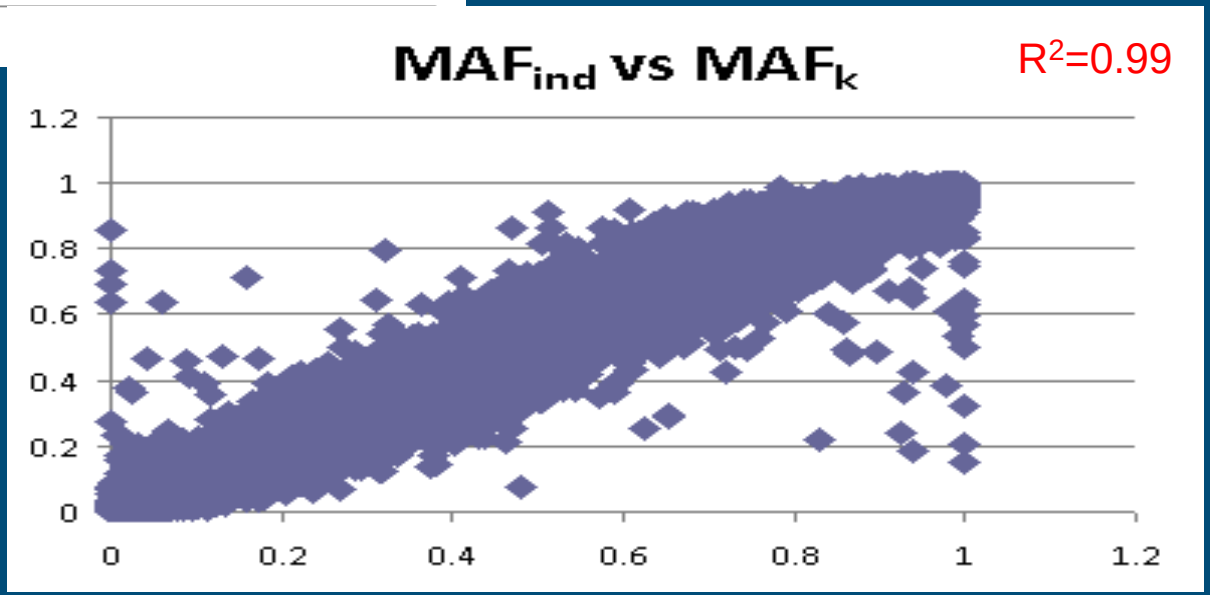
Application: Population Genetics using pools

- Pools are a cost-effective way to identify what SNPs are segregating in a specific breed
- Highly reliable estimates of allele frequencies
- Estimate Phylogeny
- Identify regions under selection (selective sweeps)

High correlation between individuals and pools



K-correction of MAF based on the signals in the heterozygote cluster



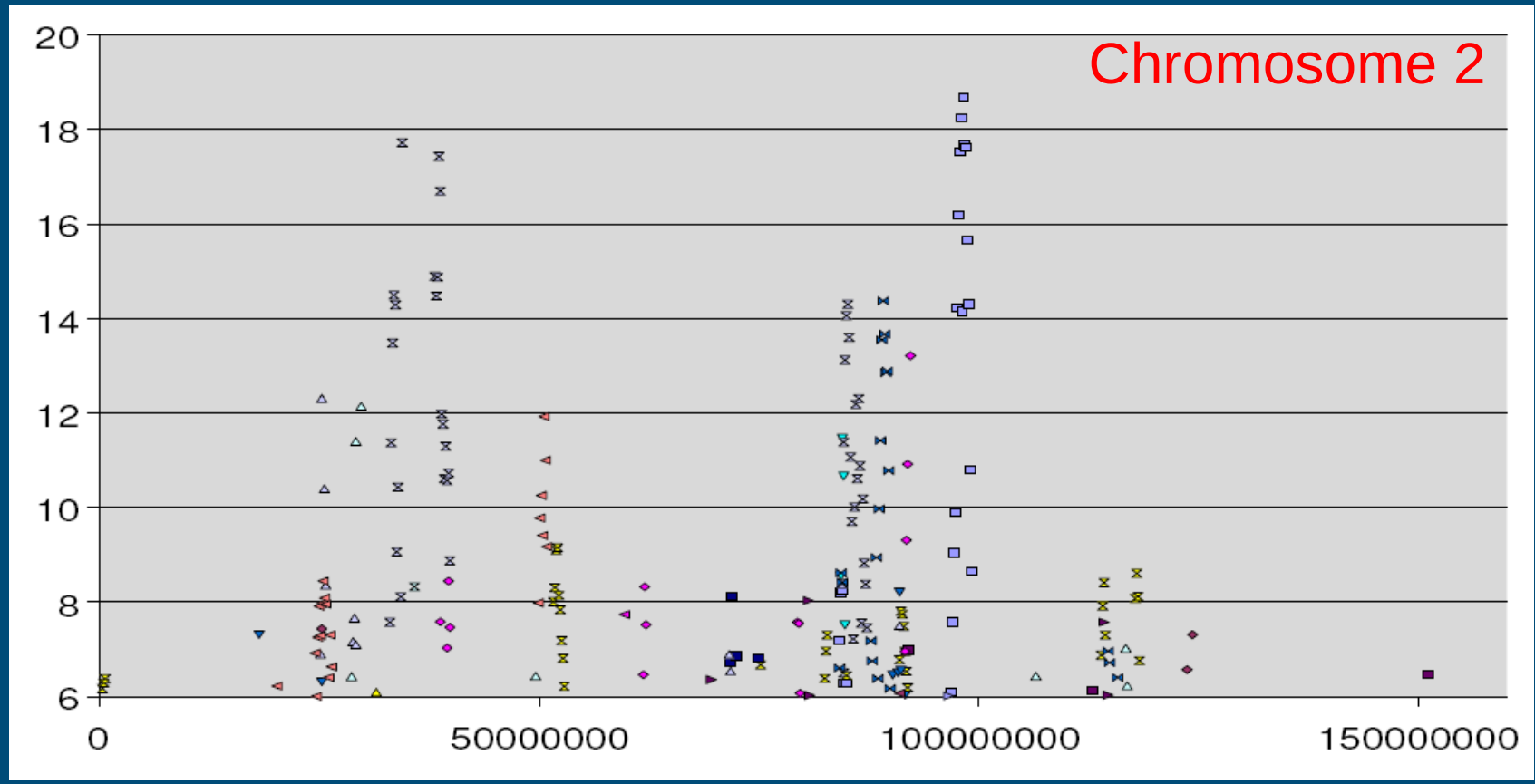
High correlation between individuals and pools

Gallus Lafayettie

Identification of the ancestral allele

Data from individually genotyped animals compared to a pooled DNA sample of that breed

Identification of selective sweeps

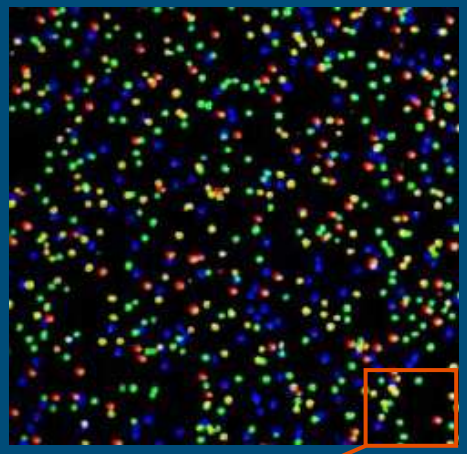


Overview

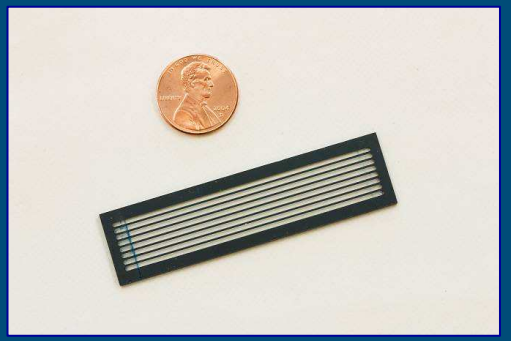
- SNP typing: Current status
- SNP discovery using 2nd generation sequencing technology (Solexa)
- Copy Number Variation (CNV)
- Re-sequencing of genomes

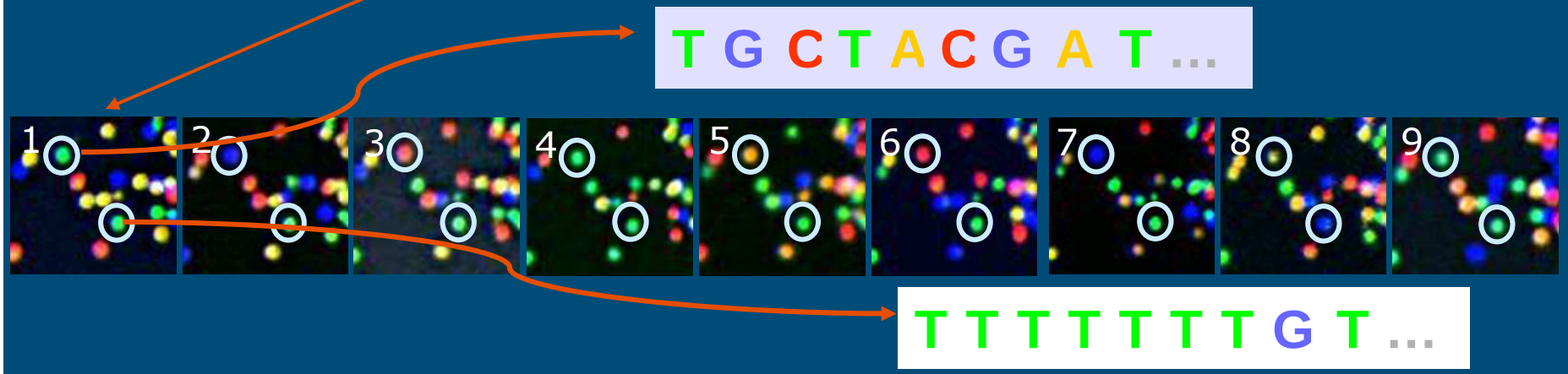
Ultrahigh throughput sequencing (Solexa, Illumina GA)

> 1 billion bp in a single run at ~€10,000



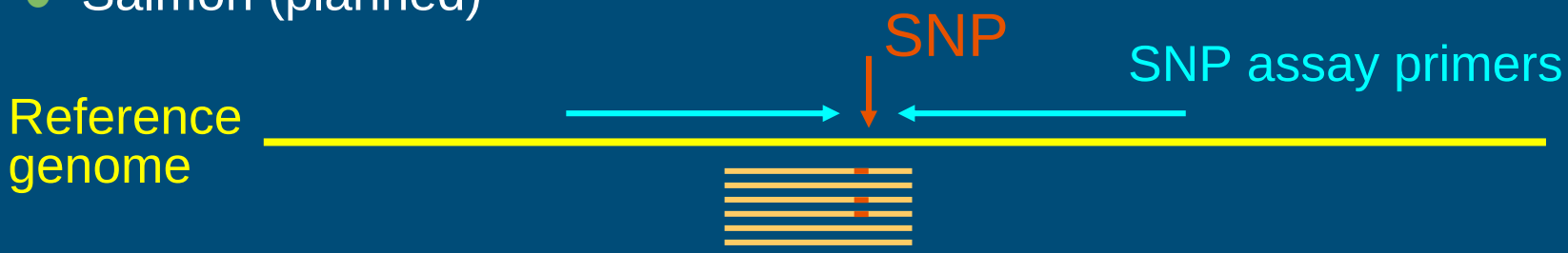
- **Read length 36 bp**
- **3-8 million reads per lane**
- **8 lanes per flow cell**
- **1 sequence run = 3 days**



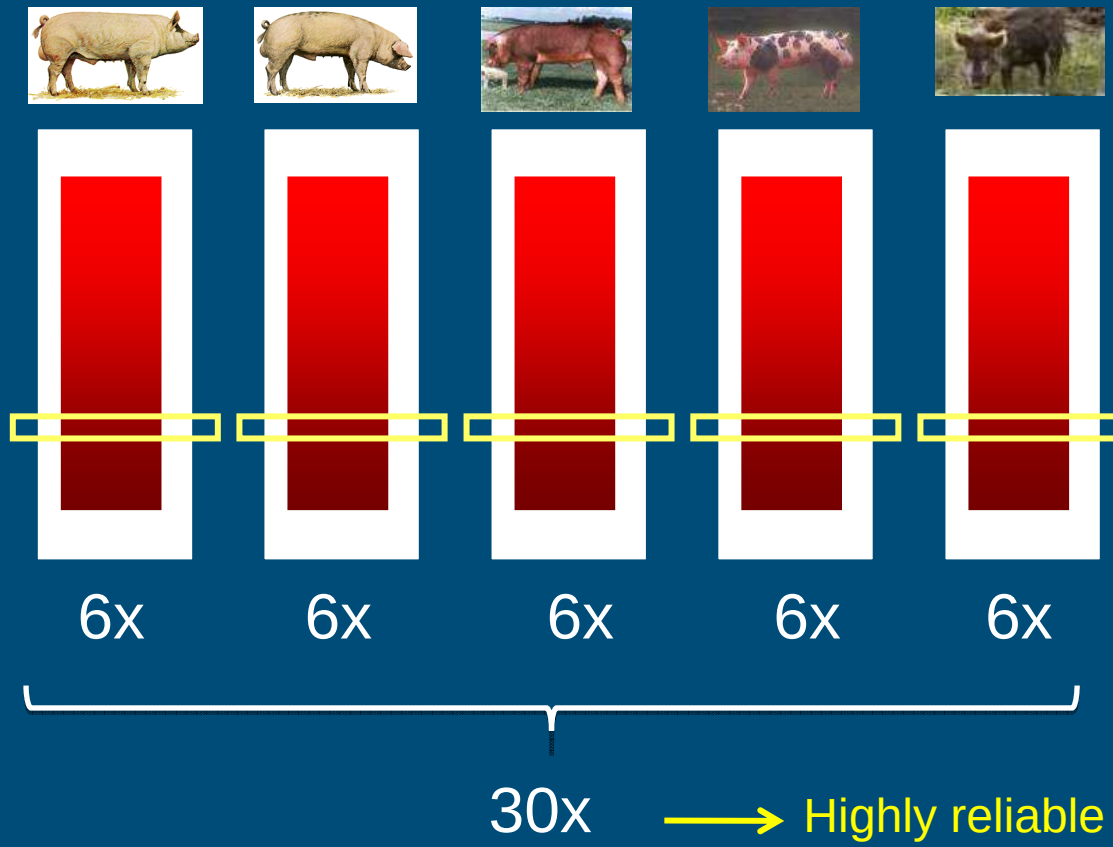


SNP detection: two strategies

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



Strategy 1: With reference genome

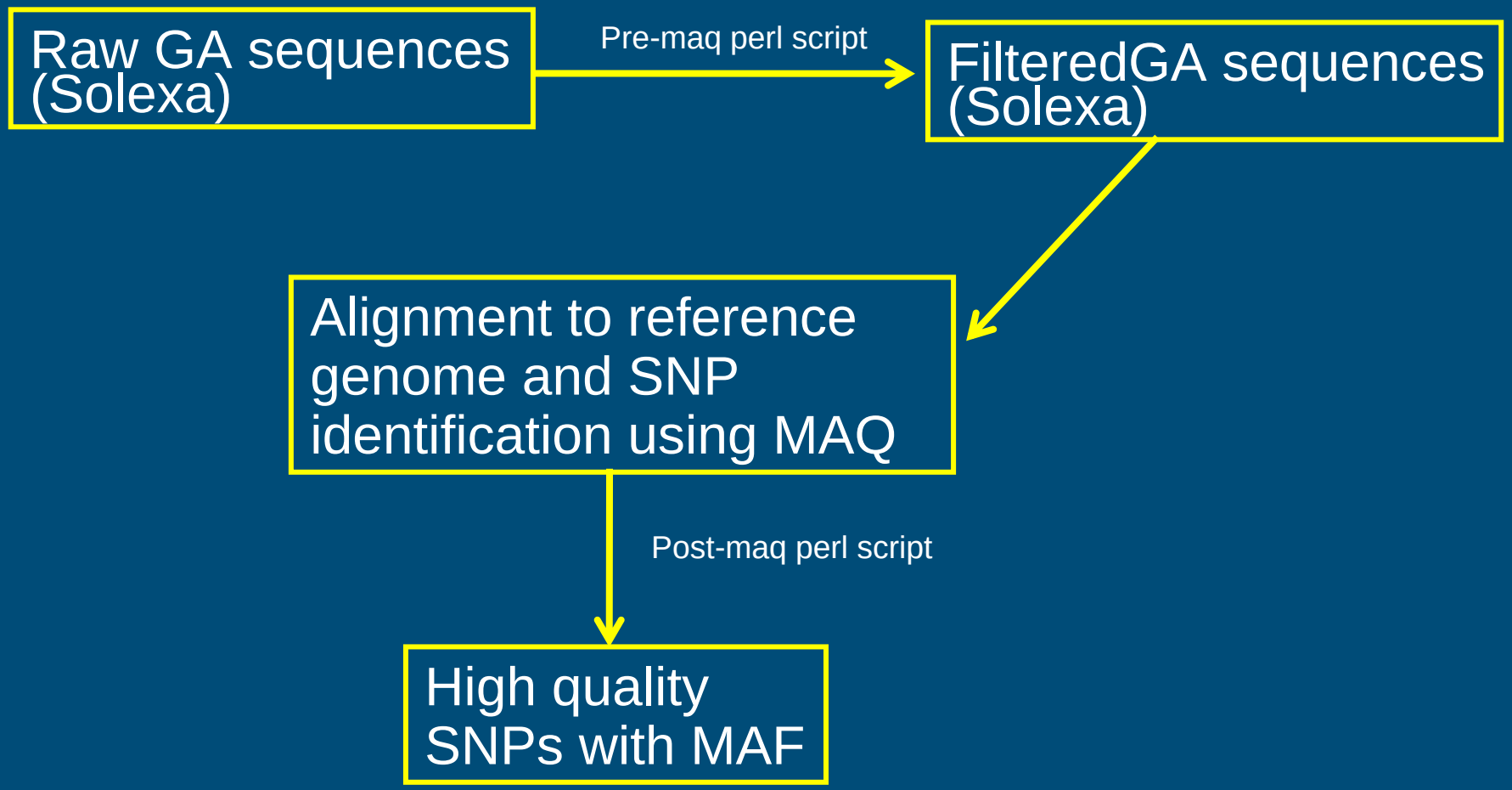


- Reduced representation library (RRL)
- Restriction digest of DNA
- Separate on acrylamide gels
- Isolate 150-200 bp fragments

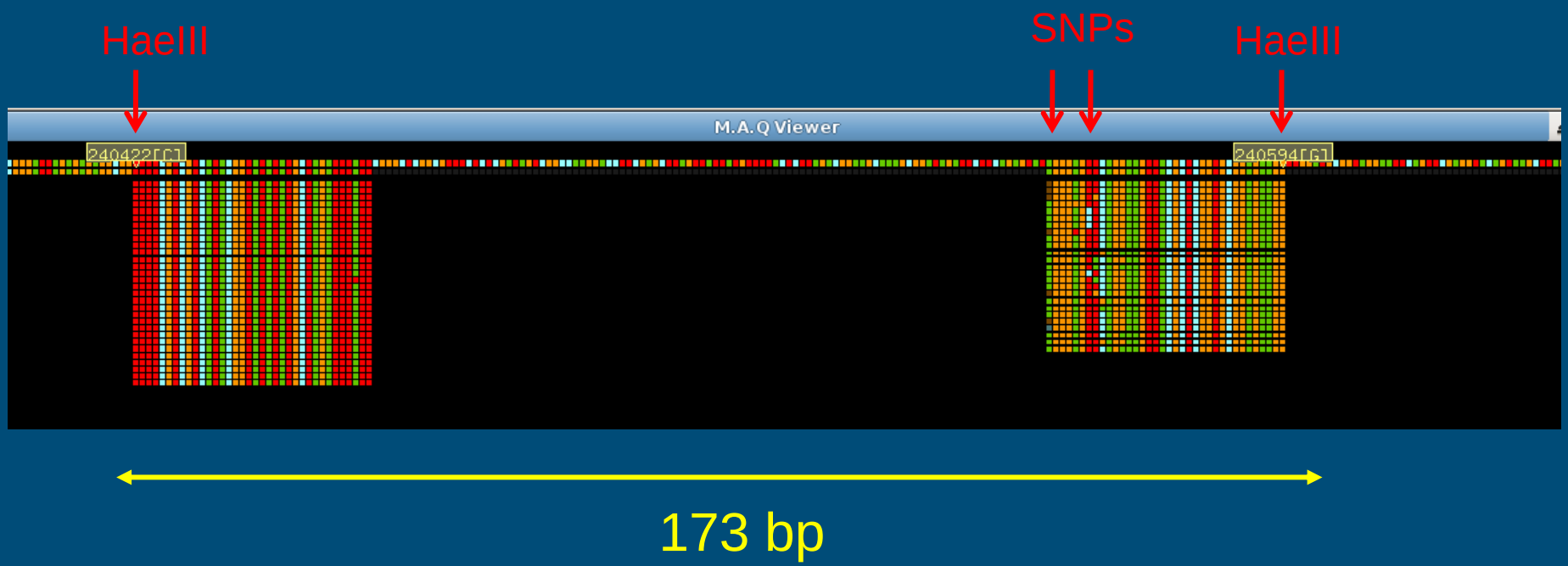
→ Estimate of MAF within breeds

→ Highly reliable SNP identification

Analysis overview



Alignment and SNP detection using MAQ



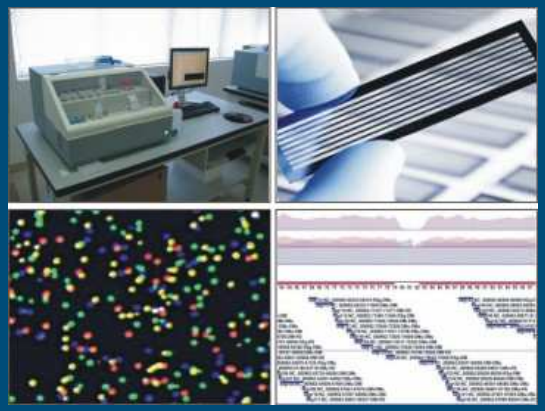
Alignment and SNP detection using MAQ

Strategy 1: With reference genome

- Pigs (finished)
  - 300 million 36 bps reads (after QC, raw data=450 million)
  - 5x5 different RRLs representing + 10% of the genome
  - 350,000 high quality SNPs MAF
  - 60K Ilumina iSelect BeadChip ordered
- Chicken
  - 115 million reads (partly analyzed)
  - 4x1 RRLs (2 Cobb, 2 HG)
  - 6 % of genome
  - detected 384,000 SNPs with MAF
  - Also cover missing chromosomes

Porcine Infinium Beadchip SNPs

- A total of 60,212 SNPs were selected
  - 41,667 with coordinates on build 7 (average spacing of 44 Kb along the genome)
  - 18,545 SNPs covering the remainder 30% of the genome
  - 4,497 of these 18,545 SNPs have location predicted by human-pig comparative mapping
  - 57,996 SNPs had minor allele frequency calculated
  - Average MAF=0.4
- Test chip analysed end of october
  - 3000 SNPs replaced
  - Final chip ready by beginning of December



The porcine HapMap project

Porcine Hapmap panel typed by Illumina

Breed	Discovery Panel	Trio's	Total #
Landrace	29	17	76
Large White	36	25	135
Duroc	34	15	79
Pietrain	23	15	95
Wild Boar	20	-	20
Hampshire	-	15	69
Berkshire	-	15	58
Meishan	-	-	30
Wild Boar	17	-	3
Other Sus. sp.	-	-	6
Tobasco			1

Sus celebensis
Sus verrucosus
Sus barbatus

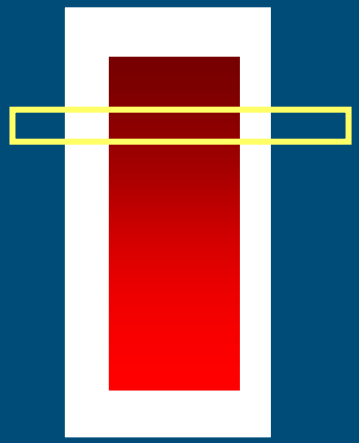
Animal being sequenced

Porcine Hapmap panel typed by WU

- Wild boar samples from all over the world
- Additional breeds from Europe
- Additional breeds from China

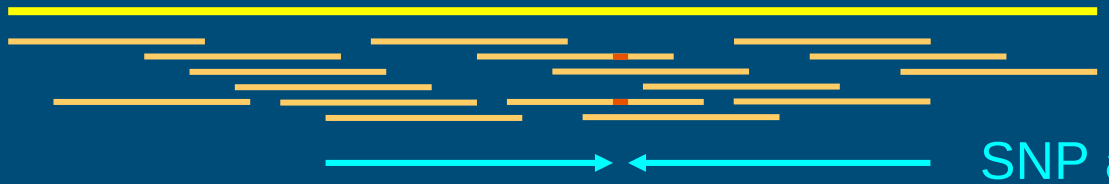
- Genotyping at Illumina: December
- Genotyping at WU: February
- Total number: ~ 1100

Strategy 2: Without reference genome

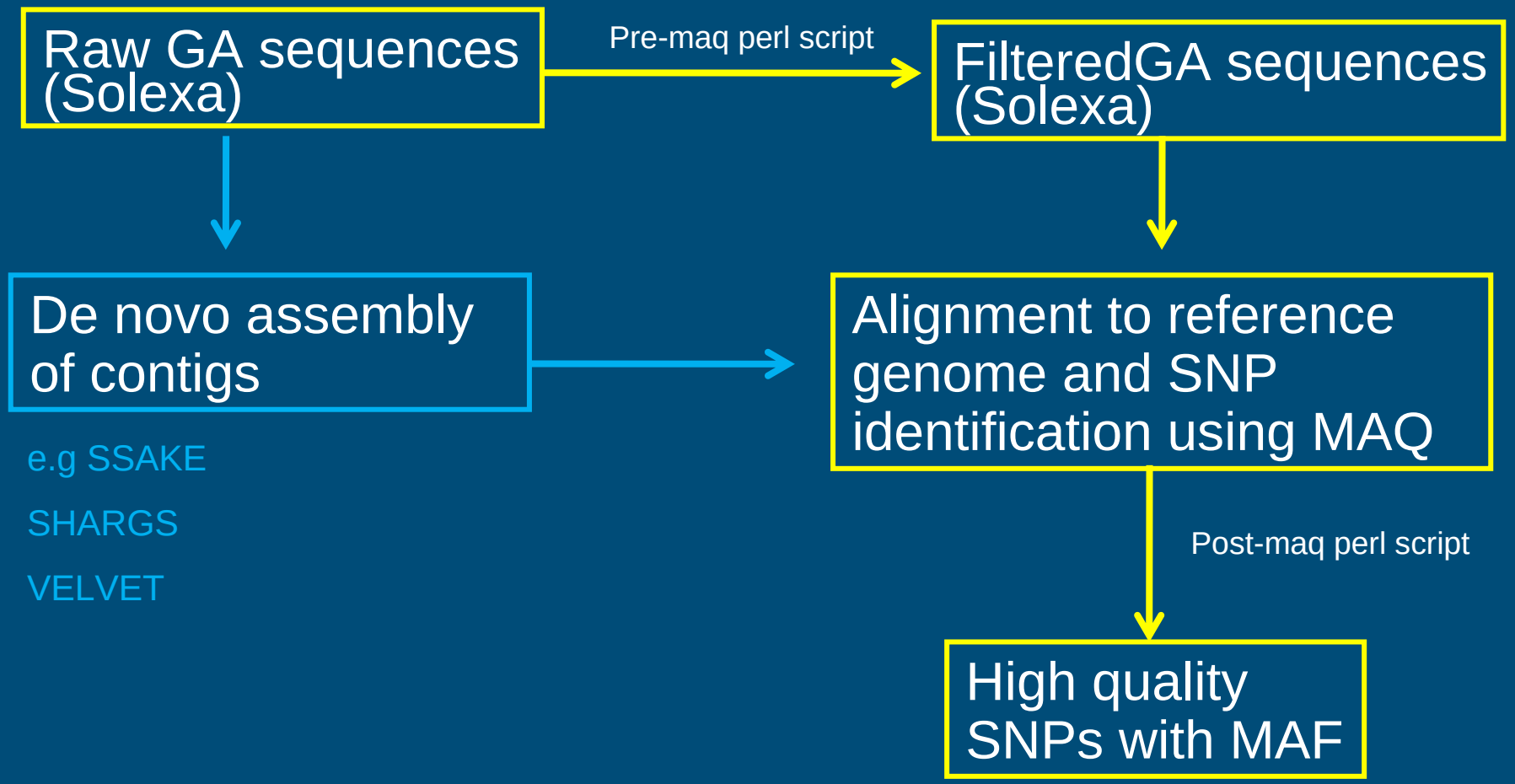


- Reduced representation library (RRL)
- Separate on agarose gel
- Isolate 2000-4000 bp fraction
- Random shearing of fragments
- Isolate 200-250 bp fragments

- Assembly of sequence contigs that can function as a reference (using a number of assemblers e.g. SSAKE, SHARGS, VELVET)



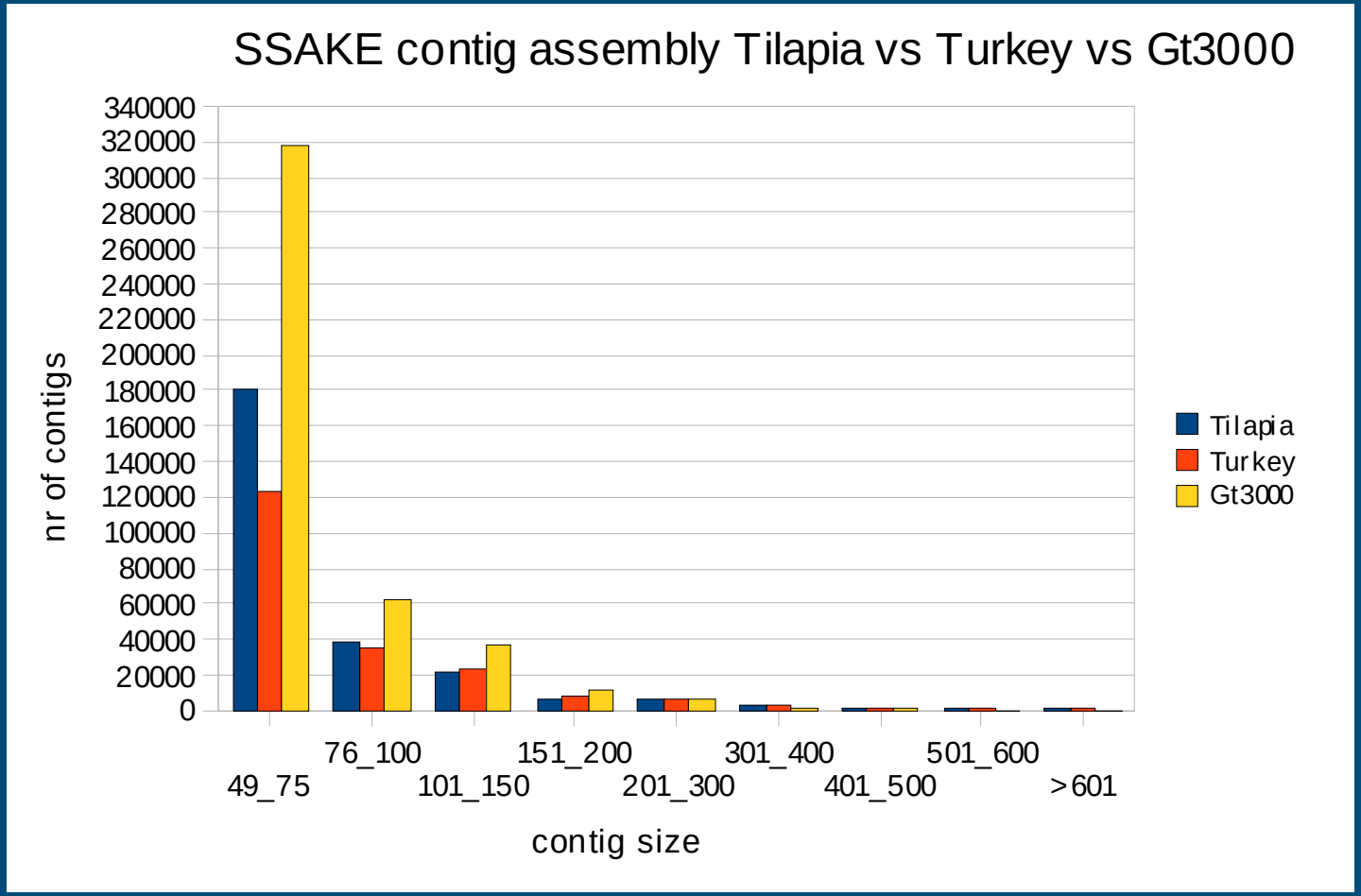
Analysis overview strategy 2



Strategy 2: Without reference genome

- Turkey
  - ~35 million 35 bp reads (after QC)
- Great Tit
  - 2 RRLs done: 50 million 36 bp (10 % of the genome)
  - 1 RRL not yet analyzed (paired end)
- Tilapia
  - 30 milion reads
- Duck
  - 30 milion reads paired end (not yet analyzed)

Size distribution of contigs



SNPs identified

- Turkey
  - 14,000 SNPs
  - 50 % have sufficient flanking sequence for assay design
  - 50 % have (unique) homology with chicken
  - 384 SNP assay ordered
- Great tit (one RRL finished)
  - 25,225 SNPs
  - Will be compared to Zebra finch genome

Overview

- SNP typing: Current status
- SNP discovery using 2nd generation sequencing technology (Solexa)
- Copy Number Variation (CNV)
- Re-sequencing of genomes

Copy number variation (CNV) in chicken

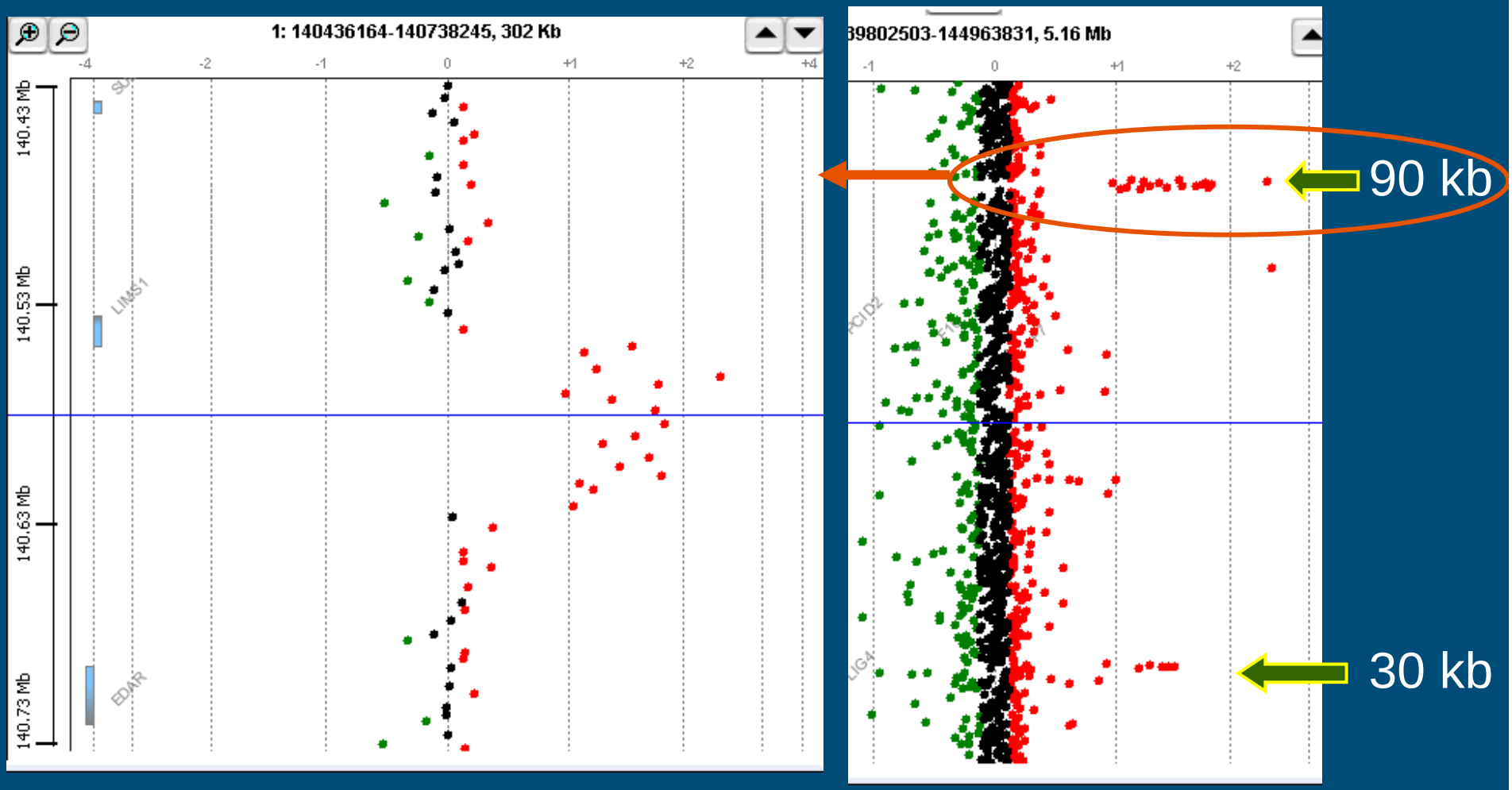
- Agilent 244K chicken array
- Analysed 80 individuals from different breeds
- Included a few small pedigrees
- Used sequenced bird (UCD001) as the reference

Chicken Genome CGH Microarray Kit 244A

High-performance with maimum precision

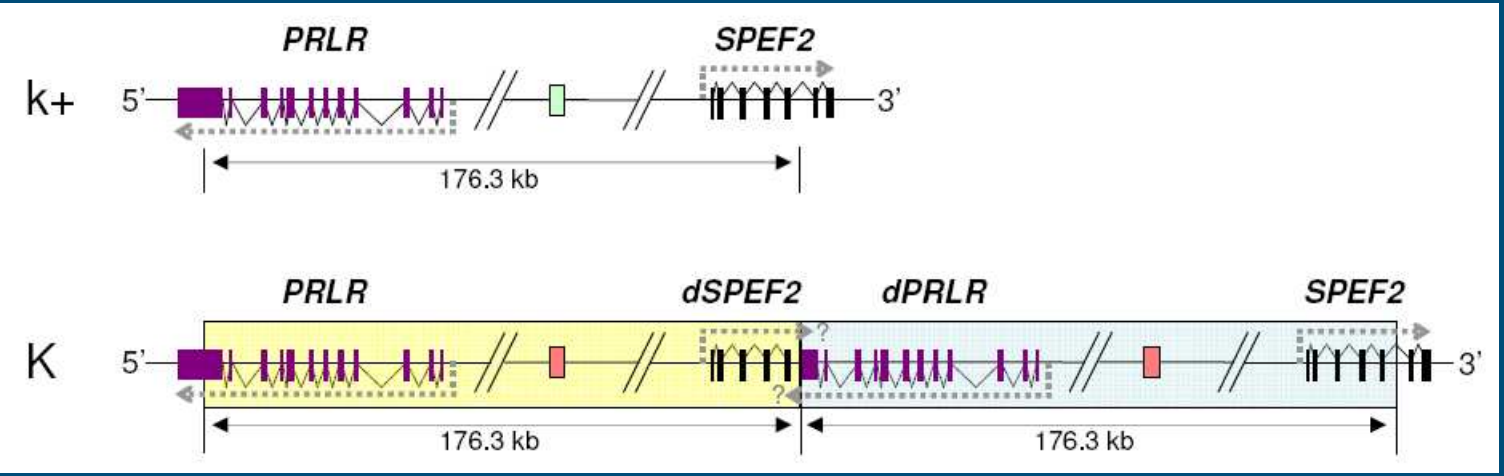
The Agilent Chicken Genome CGH Microarray Kit 244A is a high-resolution tool for genome-wide DNA copy number variation profiling without amplification or complexity reduction. Comprehensive probe coverage spans both coding and noncoding regions, promoters, miRNAs and telemeric regions. Research use only.

Example chicken chromosome 1



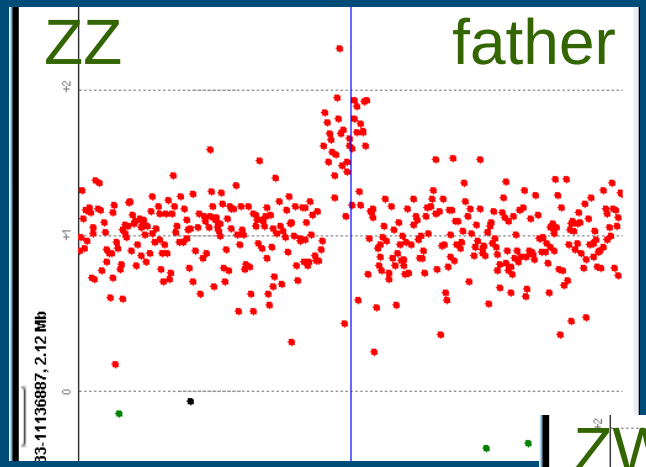
Example 2: Late feathering locus

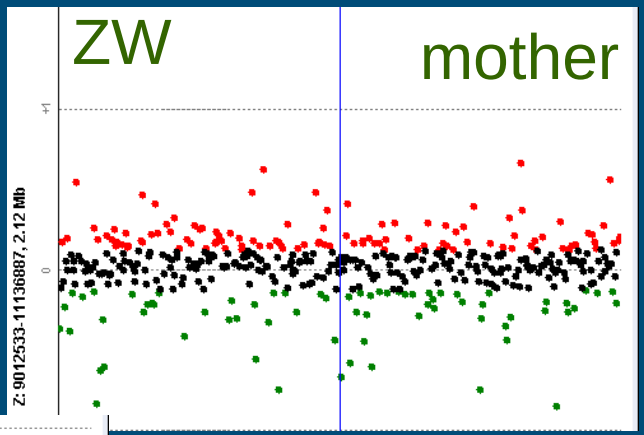
- 180 Kb duplication on Z-chromosome
  - Partial duplication of PRLR and SPEF2 gene
- Affecting feather development
- Used for sexing 1 day old chicks

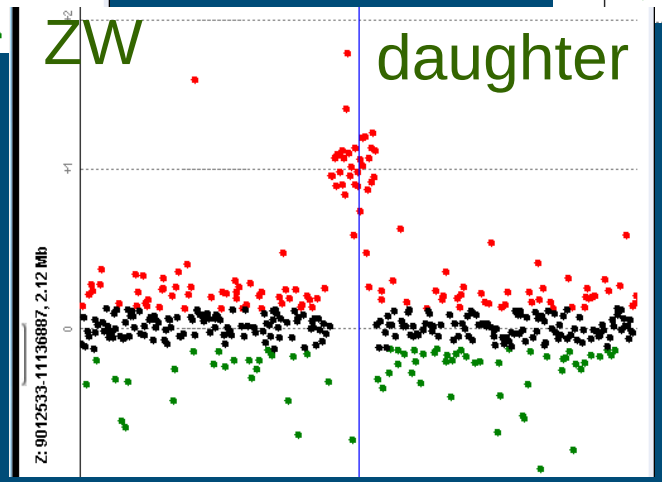


Elferink MG, Vallée AA, Jungerius AP, Crooijmans RP, Groenen MA. (2008) Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC Genomics. 9:391

Example 2: Late feathering locus







k^+ / K
3 copies

k^+ / -
1 copy

K / - 2 copies

Overview

- SNP typing: Current status
- SNP discovery using 2nd generation sequencing technology (Solexa)
- Copy Number Variation (CNV)
- Re-sequencing of genomes

Future applications in functional genomics

- Increased ultra high throughput sequencing
  - Shift from microarrays to direct sequencing of cDNA
  - Re-sequencing the complete genome of individuals
    - Identify all variations in an individual's genome
  - Sequencing to identify regulatory elements
    - E.g. miRNAs, transcription factor binding sites etc.

Published online 22 January 2008 | *Nature* **451**, 378-379 (2008) | doi:10.1038/451378b

News

International genome project launched

Three-year study will capture variation in 1,000 people.

Erika Check Hayden

A much-anticipated international project to sequence the entire genomes of 1,000 people

3rd generation sequencer with capacity to produce up to 100 billion bp of sequence

Sequencing single molecules

HeliScope™ Analysis Engine

print page | email page

As a vital component of th
the HeliScope™ Analysis
HeliScope™ Single Molecu
provides sufficient compu
real-time image analysis.
sized to support data man
operation and maximizing

HeliScope™ Single Molecule Sequencer

print page | email page

The HeliScope™ Single Molecule Sequencer is the first genetic analyzer to harness the power of direct DNA measurement, enabled by Helicos True Single Molecule Sequencing (tSMS)™ technology. As the world's first DNA microscope™, the HeliScope instrument performs tSMS chemistry and captures images to observe sequencing-by-synthesis reactions for billions of individual DNA molecules in parallel.

28 Tb of data storage (sufficient for only 2 runs !)

Acknowledgements

Chicken genome sequencing consortium

Chicken SNP consortium

Swine SNP consortium

Washington University, St. Louis

USDA, High-Density SNP Discovery, Validation and Characterization in Swine

USDA, Genome-wide marker-assisted selection over multiple generations in commercial poultry

EU-Sabre

EU-Eadgene

Thank You