64th Annual Meeting of the European Federation of Animal Science

28th August 2013

Pierre-François ROUX

Using NGS to characterize genetics of meat-type chicken lines

INRA, UMR PEGASE, Rennes, France
Animal genetics department

Biological context

- → **Challenge in genetics :** identify QTL and causative mutation underlying complex traits
- → **Purpose :** To offer tools for medicine diagnosis in human genetics or for animal selection
- → **Thousands of QTL identified**, whatever the species considered, using linkage analysis
- → But size of QTL are usually very large and contain hundreds genes : it is still **hard to identify causal mutation**

Biological context

→ **Challenge in genetics :** identify QTL and causative mutation underlying complex traits

→ **Purpose :** To offer tools for medicine diagnosis in human genetics or for animal selection

→ **Thousands of QTL identified**, whatever the species considered, using linkage analysis

→ But size of QTL are usually very large and contain hundreds genes : it is still **hard to identify causal mutation**

Technologic context

→ More and more studies based on **NGS data**

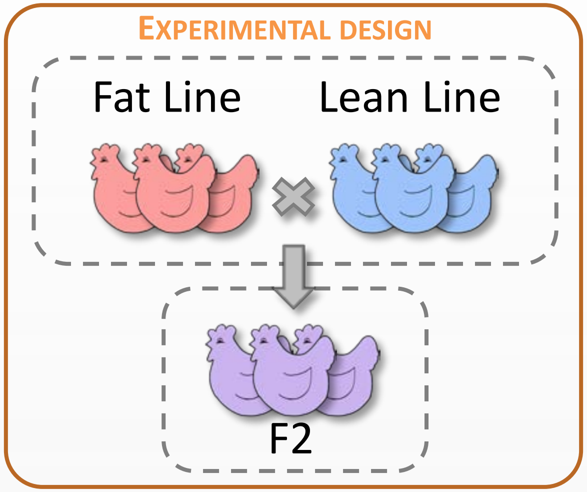
→ Some questions are still raised : impact of sequencing depth, how to filter those data …

→ Available tools for mapping, filtering, calling, annotating are numerous

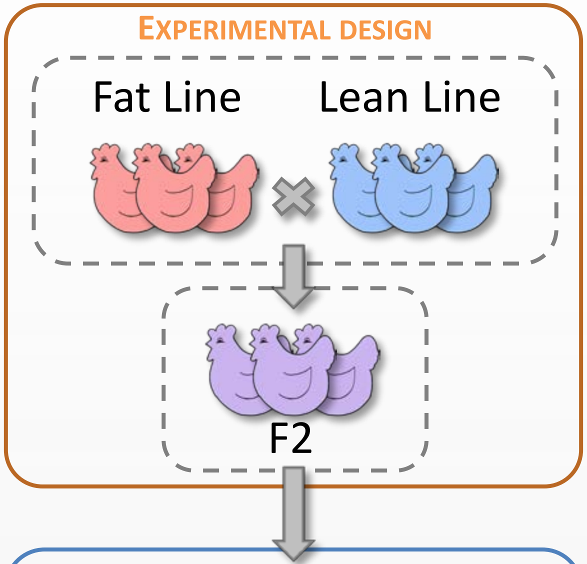
→ Interesting to study in a give study, in a given species, how reliable those data are

⇨ Use NGS to detect SNP and selective sweep in QTL region to highlight candidate mutation

Strategy



Strategy



EXPERIMENTAL DESIGN

Fat Line

Lean Line

F2

LINKAGE ANALYSIS

QTL mapping study on abdominal fatness (AF) and breast muscle weight (BMW) **Lagarrigue et al., 2006**

Strategy

DNA Sequencing
Illumina HiSeq 2000

7 F0 LL, 4 F0 FL and 9 F1 hybrid
Whole genome (20 X)
Captured 10 Mb region (100 X)

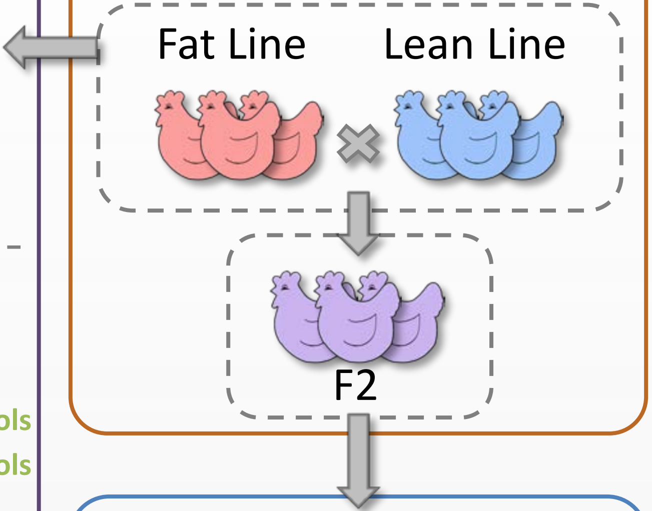
Data Processing

→Alignment on WASHUC2.1	BWA
→Filter on mapping quality	Samtools
→Removal of PCR duplicates	Samtools
→Realignment / Recalibration	GATK
→SNP calling	GATK
→Annotation	VeP

Validation step

Comparison 20 X vs 100 X	BedTools

Experimental design



Linkage Analysis

QTL mapping study on abdominal fatness (AF) and breast muscle weight (BMW) **Lagarrigue et al., 2006**

Strategy

DNA Sequencing
Illumina HiSeq 2000

7 F0 LL, 4 F0 FL and 9 F1 hybrid
Whole genome (20 X)
Captured 10 Mb region (100 X)

Data Processing

→Alignment on WASHUC2.1	BWA
→Filter on mapping quality	Samtools
→Removal of PCR duplicates	Samtools
→Realignment / Recalibration	GATK
→SNP calling	GATK
→Annotation	VeP

Validation step

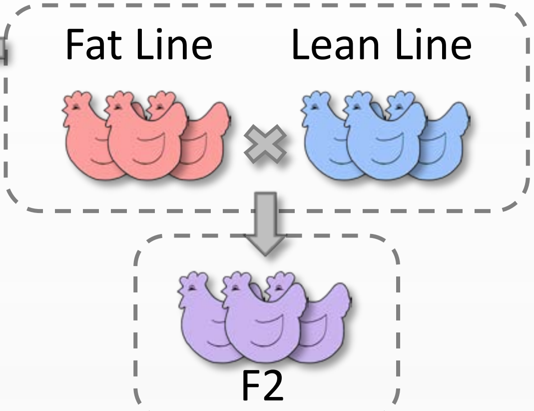
Comparison 20 X vs 100 X	BedTools

Differential Genomic Analysis

Using whole genome DNASeq data

→ Selective sweep	HapFLK

Experimental design



Linkage Analysis

QTL mapping study on abdominal fatness (AF) and breast muscle weight (BMW) **Lagarrigue et al., 2006**

Strategy

DNA SEQUENCING ILLUMINA HISEQ 2000

7 F0 LL, 4 F0 FL and 9 F1 hybrid
Whole genome (20 X)
Captured 10 Mb region (100 X)

DATA PROCESSING

→ Alignment on WASHUC2.1	BWA
→ Filter on mapping quality	Samtools
→ Removal of PCR duplicates	Samtools
→ Realignment / Recalibration	GATK
→ SNP calling	GATK
→ Annotation	VeP

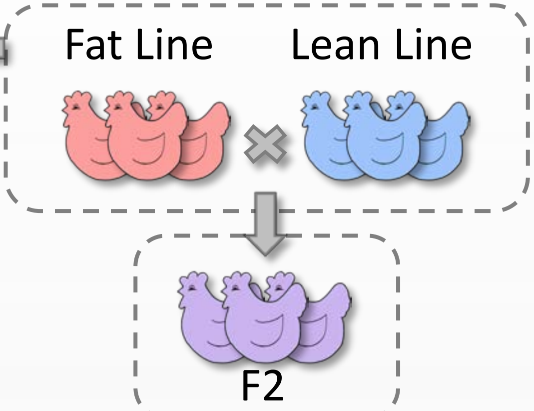
VALIDATION STEP

Comparison 20 X vs 100 X BedTools

DIFFERENTIAL GENOMIC ANALYSIS

Using whole genome DNASeq data
→ Selective sweep HapFLK

EXPERIMENTAL DESIGN



LINKAGE ANALYSIS

QTL mapping study on abdominal fatness (AF) and breast muscle weight (BMW) **Lagarrigue et al., 2006**

INTEGRATION PHASE

→ **Reduce QTL size** with sweeps
→ **Identify potential candidate genes and causal mutations** using DNASeq

Results

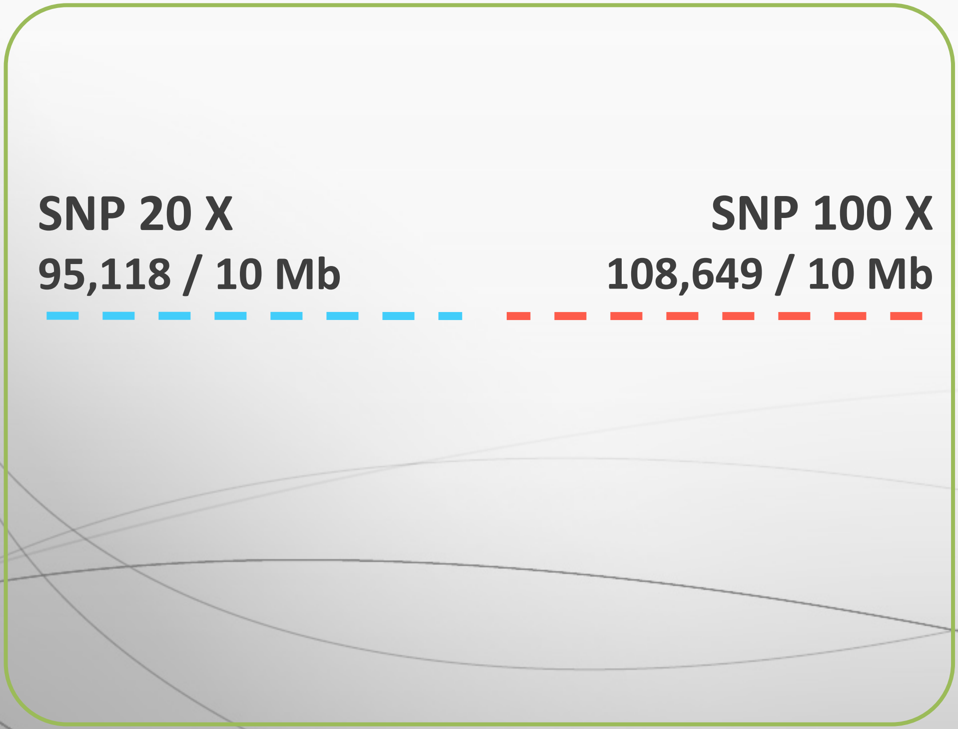
COMPARISON BETWEEN 20 X AND 100 X SNP CALLING RESULTS

Results

Comparison between 20 X and 100 X SNP calling results

➜ **Context**

Context



SNP 20 X

95,118 / 10 Mb

SNP 100 X

108,649 / 10 Mb

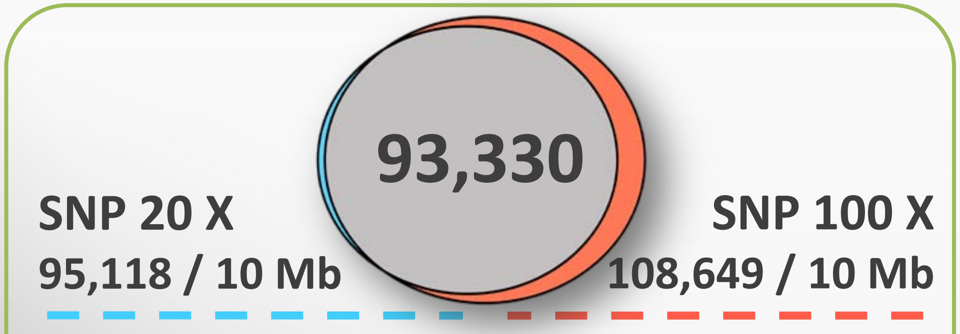
⇨ Same individuals sequenced with both 20 X and 100 X on a 10 Mb window

Results

COMPARISON BETWEEN 20 X AND 100 X SNP CALLING RESULTS

- ➜ **Context**
- ➜ **Intersection** between the two sets of SNP

Intersection



⇨ Same individuals sequenced with both 20 X and 100 X on a 10 Mb window

Results

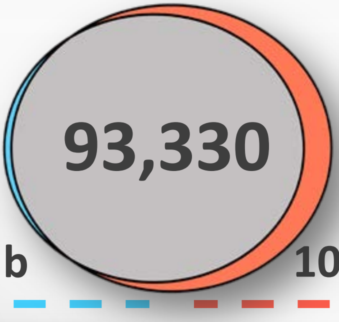
COMPARISON BETWEEN 20 X AND 100 X SNP CALLING RESULTS

➜ **Context**

➜ **Intersection** between the two sets of SNP

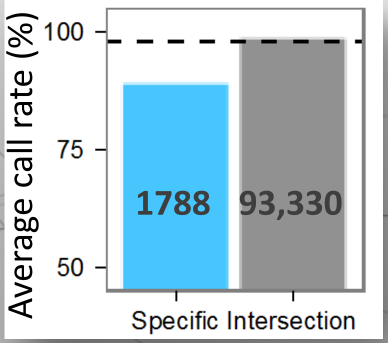
➜ **Characteristics** of 20 X specific SNP

20 X specific SNP



SNP 20 X
95,118 / 10 Mb

SNP 100 X
108,649 / 10 Mb



⇨ Same individuals sequenced with both 20 X and 100 X on a 10 Mb window

⇨ **SNP 20 X :** 98 % in intersection
2 % specific

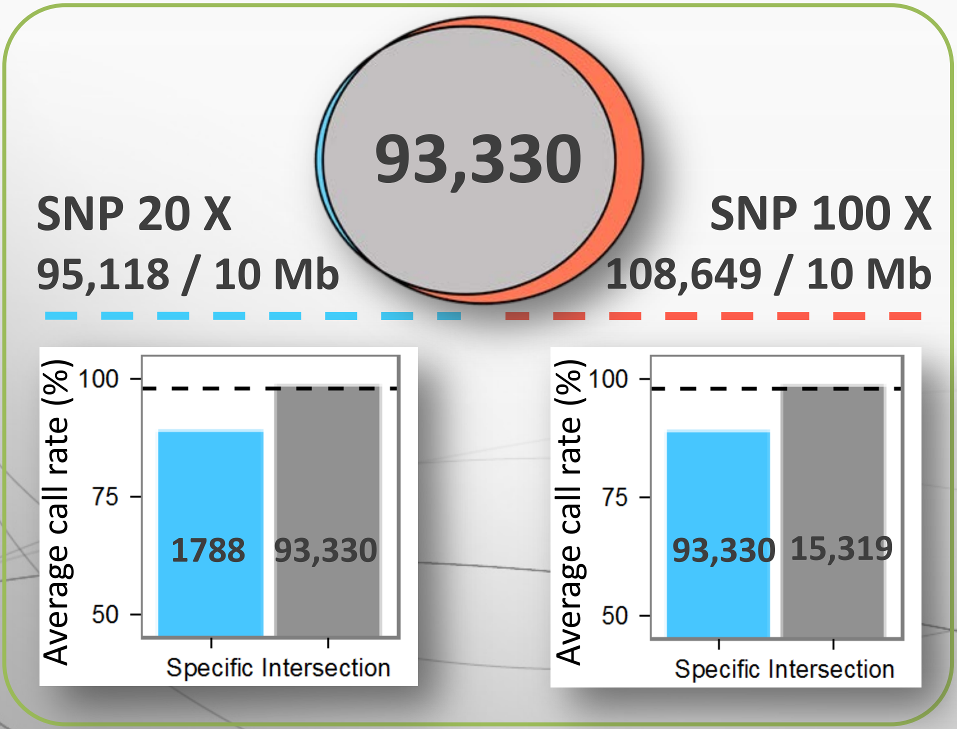
But specific SNP are, in average, of poor quality (average call rate < 90 %) and must be considered as false positives

Results

COMPARISON BETWEEN 20 X AND 100 X SNP CALLING RESULTS

- ➜ **Context**
- ➜ **Intersection** between the two sets of SNP
- ➜ **Characteristics** of 20 X specific SNP
- ➜ **Characteristics** of 100 X specific SNP

100 X specific SNP



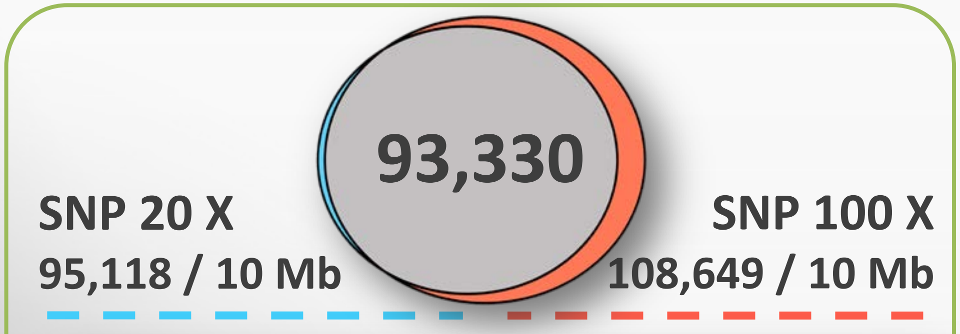
- ⇨ Same individuals sequenced with both 20 X and 100 X on a 10 Mb window
- ⇨ **SNP 20 X :** 98 % in intersection
  2 % specific
  But specific SNP are, in average, of poor quality (average call rate < 90 %) and must be considered as false positives
- ⇨ **SNP 100 X :** 86 % in intersection
  14 % specific
  Specific SNP have an average call rate of 92 % i.e. among them there are false positives and reliable SNP

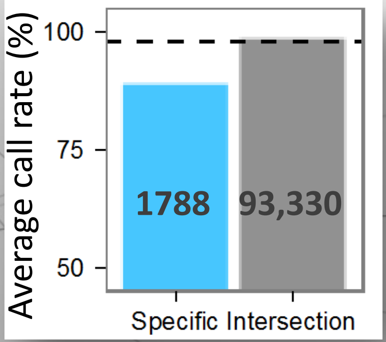
Results

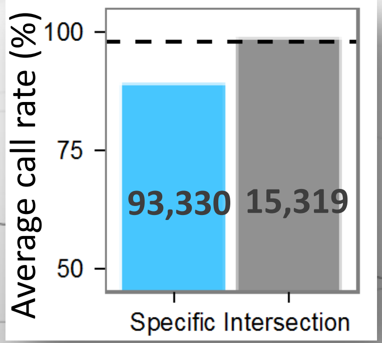
COMPARISON BETWEEN 20 X AND 100 X SNP CALLING RESULTS

- ➜ **Context**
- ➜ **Intersection** between the two sets of SNP
- ➜ **Characteristics** of 20 X specific SNP
- ➜ **Characteristics** of 100 X specific SNP

Conclusion







- ⇨ Same individuals sequenced with both 20 X and 100 X on a 10 Mb window
- ⇨ **SNP 20 X :** 98 % in intersection
  2 % specific
  But specific SNP are, in average, of poor quality (average call rate < 90 %) and must be considered as false positives
- ⇨ **SNP 100 X :** 86 % in intersection
  14 % specific
  Specific SNP have an average call rate of 92 % i.e. among them there are false positives and reliable SNP
- ⇨ 20 X re-sequencing is sufficient to have information for almost all SNP

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

➜ General characteristics

General characteristics

Chicken genome size :	1.05 Gb
SNP per individual :	2.7 M (± 0.5)
SNP density :	2.6 / kb
SNP in the global scheme :	9.4 M

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

➜ **General characteristics**

General characteristics

Chicken genome size :	1.05 Gb
SNP per individual :	2.7 M (± 0.5)
SNP density :	2.6 / kb
SNP in the global scheme :	9.4 M

⇨ SNPs density in accordance with literature

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

➜ **General characteristics**

➜ **Genome wide functional consequences :** global overview

Functional consequences

	SNP		Genes
All types of SNP	**N**	**%**	**N**
Intergenic	4513004	48,68	-
Regulatory regions	776080	8,37	17352
Intronic	3873604	41,78	14087
Coding - Synonymous	65276	0,70	11905
Coding - Non Synonymous	43410	0,47	10676

⇨ SNPs density in accordance with literature

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

➜ **General characteristics**

➜ **Genome wide functional consequences :** global overview

Functional consequences

	SNP		Genes
All types of SNP	**N**	**%**	**N**
Intergenic	4513004	48,68	-
Regulatory regions	776080	8,37	17352
Intronic	3873604	41,78	14087
Coding - Synonymous	65276	0,70	11905
Coding - Non Synonymous	43410	0,47	10676

⇨ SNPs density in accordance with literature

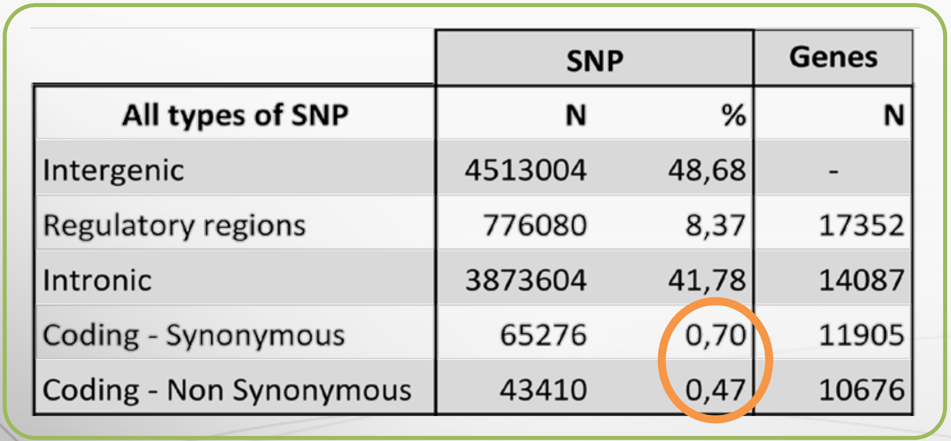
⇨ Approximately 50 % of SNP are in inter-genic regions

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

- ➜ **General characteristics**
- ➜ **Genome wide functional consequences :** global overview

Functional consequences



	SNP		Genes
All types of SNP	**N**	**%**	**N**
Intergenic	4513004	48,68	-
Regulatory regions	776080	8,37	17352
Intronic	3873604	41,78	14087
Coding - Synonymous	65276	0,70	11905
Coding - Non Synonymous	43410	0,47	10676

- ⇨ SNPs density in accordance with literature
- ⇨ Approximately 50 % of SNP are in inter-genic regions
- ⇨ Only 1.2 % are located on genes

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

- ➜ **General characteristics**
- ➜ **Genome wide functional consequences :** global overview

Functional consequences

	SNP		Genes
All types of SNP	**N**	**%**	**N**
Intergenic	4513004	48,68	-
Regulatory regions	776080	8,37	17352
Intronic	3873604	41,78	14087
Coding - Synonymous	65276	0,70	11905
Coding - Non Synonymous	43410	0,47	10676

- ⇨ SNPs density in accordance with literature
- ⇨ Approximately 50 % of SNP are in inter-genic regions
- ⇨ Only 1.2 % are located on genes
- ⇨ Among the 17,838 genes listed on the chicken genome, 60 % have at list 1 non synonymous SNP

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

- **General characteristics**
- **Genome wide functional consequences :** global overview
- **Genome wide functional consequences :** focus on coding SNP

Functional consequences

	SNP		Genes
Coding SNP	**N**	**%**	**N**
Synonymous	65276	59,45	11905
Missense	27711	25,24	8214
Initiator or stop codon	495	0,45	461
Splicing site	16286	14,83	7155
Mature mi-RNA	34	0,03	30

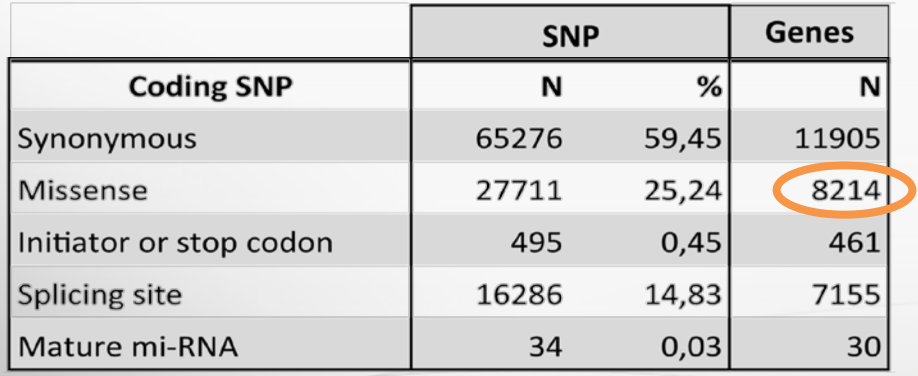
- SNPs density in accordance with literature
- Approximately 50 % of SNP are in inter-genic regions
- Only 1.2 % are located on genes
- Among the 17,838 genes listed on the chicken genome, 60 % have at list 1 non synonymous SNP
- 60 % of the SNP located on genes are synonymous

Results

CHARACTERIZATION OF WHOLE GENOME SNP IN OUR LINES

- ➜ **General characteristics**
- ➜ **Genome wide functional consequences :** global overview
- ➜ **Genome wide functional consequences :** focus on coding SNP

Functional consequences



	SNP		**Genes**
Coding SNP	**N**	**%**	**N**
Synonymous	65276	59,45	11905
Missense	27711	25,24	8214
Initiator or stop codon	495	0,45	461
Splicing site	16286	14,83	7155
Mature mi-RNA	34	0,03	30

- ⇨ SNPs density in accordance with literature
- ⇨ Approximately 50 % of SNP are in inter-genic regions
- ⇨ Only 1.2 % are located on genes
- ⇨ Among the 17,838 genes listed on the chicken genome, 60 % have at list 1 non synonymous SNP
- ⇨ 60 % of the SNP located on genes are synonymous
- ⇨ Missense SNP concern 45 % of genes

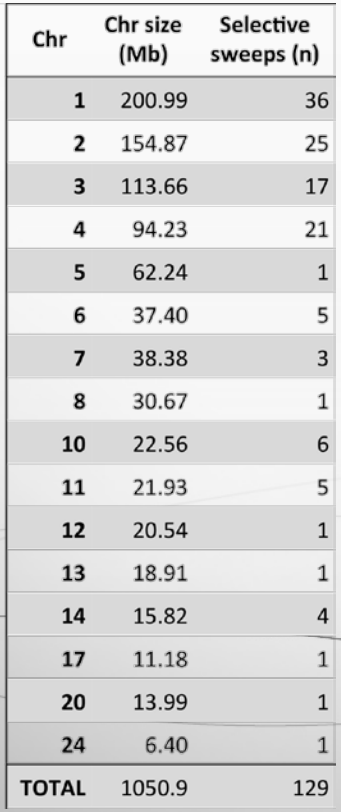
Results

GENOME SCAN TO DETECT SELECTIVE SWEEPS USING HAPFLK

Results

Genome scan to detect selective sweeps using HapFLK

Selective sweeps



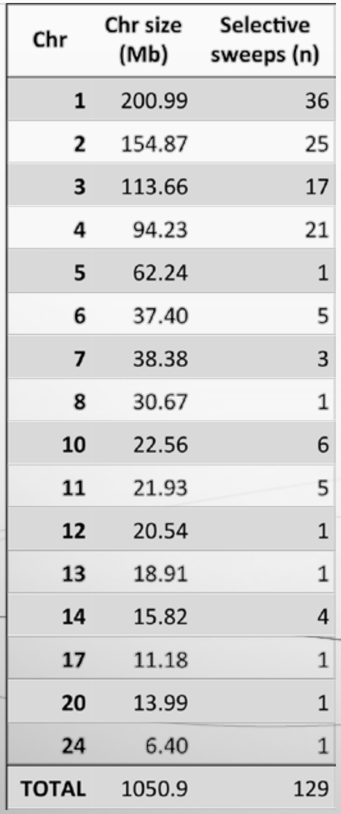
Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

Results

Genome scan to detect selective sweeps using HapFLK

➜ Number of sweeps

Selective sweeps



Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

⇨ **129 sweeps** distributed on **16 chromosomes**

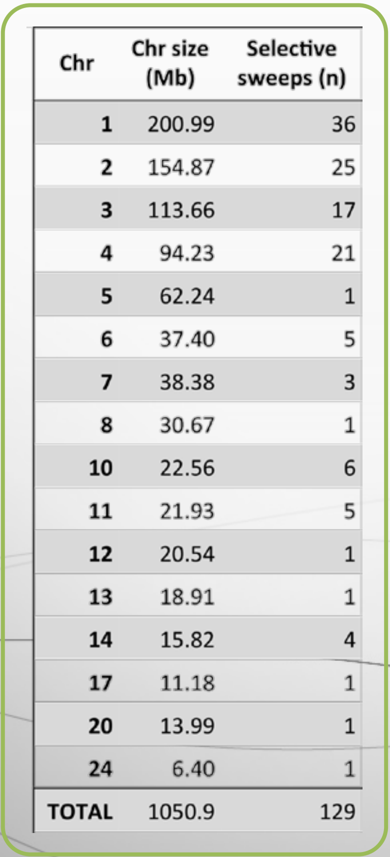
Results

Genome scan to detect selective sweeps using HapFLK

➜ Number of sweeps

➜ Size of sweeps

Selective sweeps



Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

⇨ **129 sweeps** distributed on **16 chromosomes**

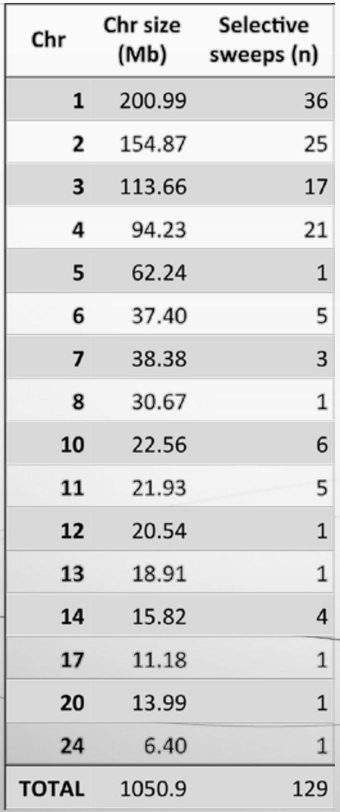
⇨ **Average size :** 98 kb (± 90)

Results

GENOME SCAN TO DETECT SELECTIVE SWEEPS USING HAPFLK

- ➜ Number of sweeps
- ➜ Size of sweeps
- ➜ Number of SNP per sweep

Selective sweeps



Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

- ⇨ **129 sweeps** distributed on **16 chromosomes**
- ⇨ **Average size :** 98 kb (± 90)
- ⇨ **Average number of SNP in sweeps :** 850 (± 700)

Results

Genome scan to detect selective sweeps using HapFLK

➜ **Number of sweeps**
➜ **Size of sweeps**
➜ **Number of SNP per sweep**
➜ **Number of genes per sweep**

Selective sweeps

Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

⇨ **129 sweeps** distributed on **16 chromosomes**

⇨ **Average size :** 98 kb (± 90)

⇨ **Average number of SNP in sweeps :** 850 (± 700)

⇨ **Average number of genes in sweeps :** 2.2 (± 1.5)

Results

OVERLAYING QTL MAPPING RESULTS WITH SELECTIVE SWEEP ANALYSIS RESULTS

Results

OVERLAYING QTL MAPPING RESULTS WITH SELECTIVE SWEEP ANALYSIS RESULTS

➜ **QTL mapping results**

⇨ 7 QTL considered in the analysis (6 for AF, 1 for BMW)

Results

OVERLAYING QTL MAPPING RESULTS WITH SELECTIVE SWEEP ANALYSIS RESULTS

➜ **QTL mapping results**

➜ **Overlaying with HapFLK results**

⇨ 7 QTL considered in the analysis (6 for AF, 1 for BMW)

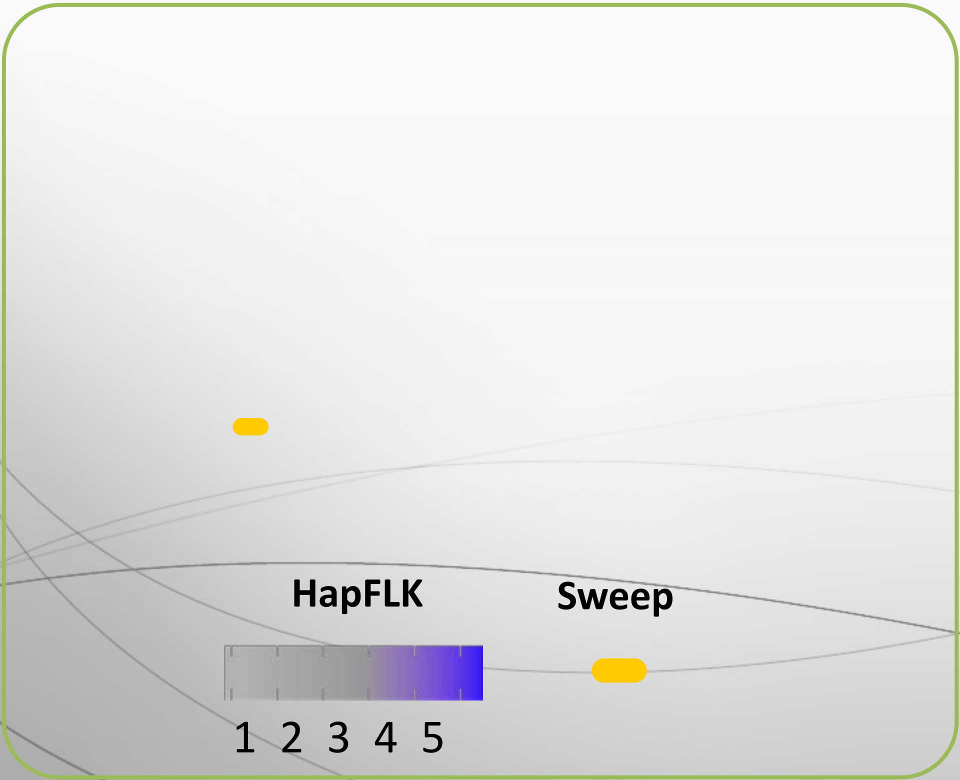
⇨ All QTL have at least one selective sweep

Results

OVERLAYING QTL MAPPING RESULTS WITH SELECTIVE SWEEP ANALYSIS RESULTS

- ➜ **QTL mapping results**
- ➜ **Overlaying with HapFLK results**
- ➜ **Examples**

Examples of co-location between QTL and sweeps

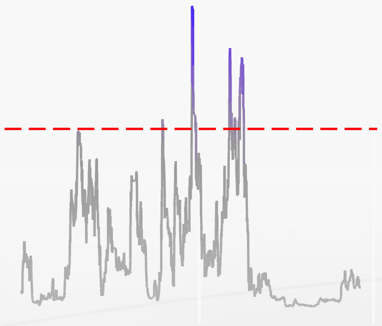


- ⇨ 7 QTL considered in the analysis (6 for AF, 1 for BMW)
- ⇨ All QTL have at least one selective sweep
- ⇨ 4 QTL have only one selective sweep

Results

OVERLAYING QTL MAPPING RESULTS WITH SELECTIVE SWEEP ANALYSIS RESULTS

- ➜ **QTL mapping results**
- ➜ **Overlaying with HapFLK results**
- ➜ **Examples**



- ⇨ 7 QTL considered in the analysis (6 for AF, 1 for BMW)
- ⇨ All QTL have at least one selective sweep

Conclusion

⇨ **20 X** sequencing depth is sufficient to study **most of SNP** and to give a highly reliable information after filtering steps

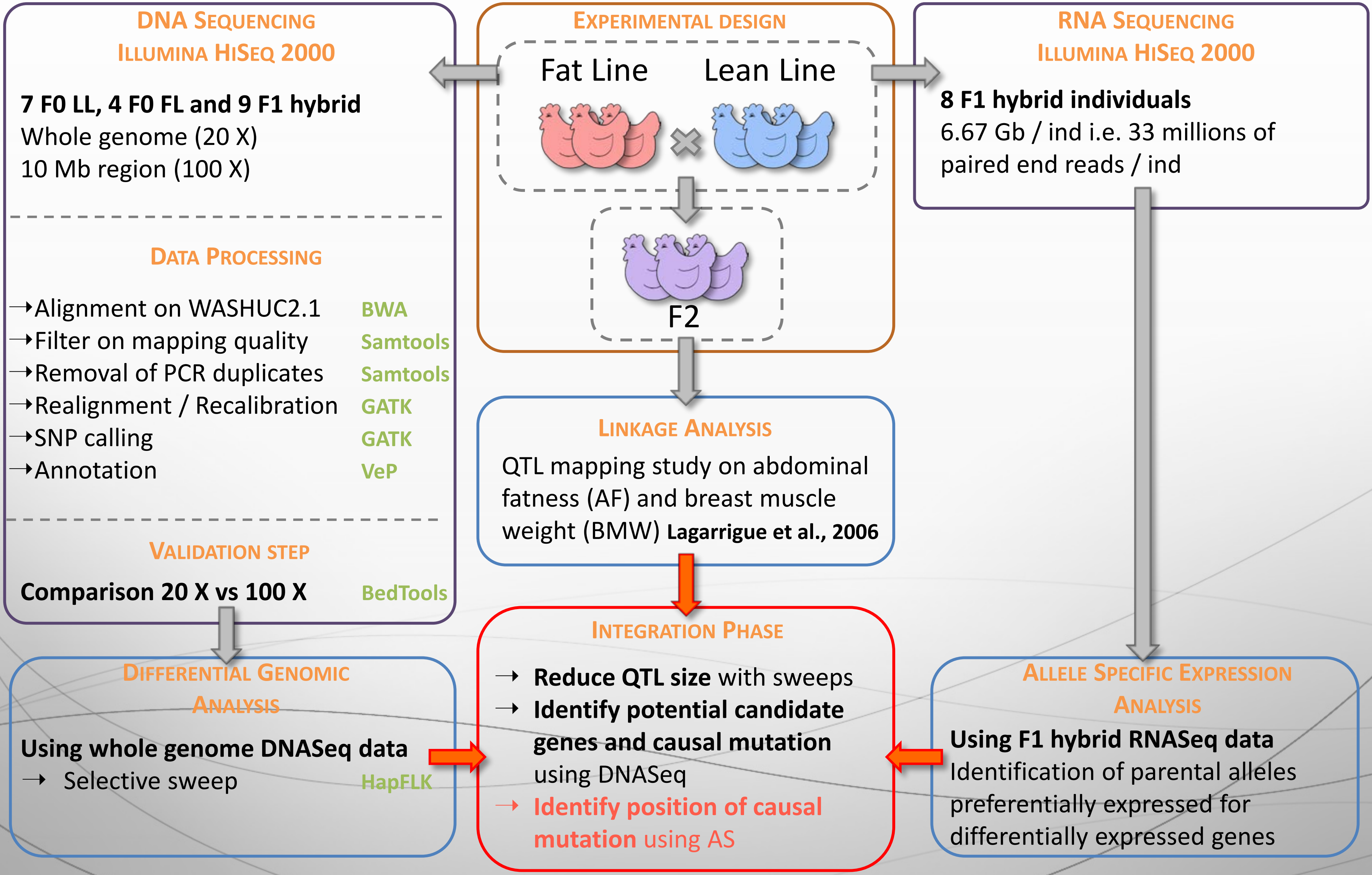
⇨ **HapFLK**, adapted for **analyses on re-sequencing data**, and to take into account genotypes of F1 individuals, allowed us to identify numerous selective sweeps along chicken genome

⇨ The overlaying step between **QTL** and **selective sweep** results allowed a **reduction of the size of the region to focus on**

⇨ On these reduced QTL regions we finally identified some interesting **functional candidate genes**

⇨ NGS data also allowed identification of **candidate mutations**

Perspective



DNA Sequencing
Illumina HiSeq 2000

7 F0 LL, 4 F0 FL and 9 F1 hybrid
Whole genome (20 X)
10 Mb region (100 X)

Data Processing

→Alignment on WASHUC2.1	BWA
→Filter on mapping quality	Samtools
→Removal of PCR duplicates	Samtools
→Realignment / Recalibration	GATK
→SNP calling	GATK
→Annotation	VeP

Validation step

Comparison 20 X vs 100 X BedTools

Experimental design

Fat Line Lean Line

F2

RNA Sequencing
Illumina HiSeq 2000

8 F1 hybrid individuals
6.67 Gb / ind i.e. 33 millions of paired end reads / ind

Linkage Analysis

QTL mapping study on abdominal fatness (AF) and breast muscle weight (BMW) **Lagarrigue et al., 2006**

Differential Genomic Analysis

Using whole genome DNASeq data
→ Selective sweep HapFLK

Integration Phase

- → **Reduce QTL size** with sweeps
- → **Identify potential candidate genes and causal mutation** using DNASeq
- → **Identify position of causal mutation** using AS

Allele Specific Expression Analysis

Using F1 hybrid RNASeq data
Identification of parental alleles preferentially expressed for differentially expressed genes

64th Annual Meeting of the European Federation of Animal Science

- **INRA, UMR PEGASE, Rennes, France Genetics & Genomic Team**
  - Sandrine Lagarrigue
  - Olivier Demeure
  - Colette Désert
  - Frederic Lecerf
- **INRA, LGC, Toulouse, France**
  - Simon Boitard
  - Bertrand Servin
  - Frédérique Pitel
- **INRA, SIGENAE, Toulouse, France**
  - Anis Djari
- **INRA, Genotoul Plate-form, Toulouse, France**
  - Diane Esquerre
- **INRA, UMR GABI, CRB GADIE, Jouy-en-Josas, France**
  - Sylvain Marthey
  - Marco Moroldo
- **INRA, UMR GABI, Jouy-en-Josas, France**
  - Tatiana Zerjal
  - Jordi Estelle
  - Bertrand Bed'hom
  - Michelle Tixier - Boichard
- **INRA, UR 38, Avian research, Tours, France**
  - Elisabeth Le Bihan - Duval

28th August 2013

Pierre-François ROUX

Many thanks for you attention