



HAL
open science

Combining Genetics Mapping, Selective Sweeps, Genome Re-Sequencing and Allele Specific Expression Reveals Several Serious Causative Genes for QTL Regions

Pierre-François Roux, Simon Boitard, Anis Djari, Diane Esquerre, Colette Désert, Morgane Boutin, Sylvain Marthey, Frédéric Lecerf, Elisabeth Duval, Christophe Klopp, et al.

► **To cite this version:**

Pierre-François Roux, Simon Boitard, Anis Djari, Diane Esquerre, Colette Désert, et al.. Combining Genetics Mapping, Selective Sweeps, Genome Re-Sequencing and Allele Specific Expression Reveals Several Serious Causative Genes for QTL Regions. International Plant & Animal Genome XXII (PAG), Jan 2014, San Diego, United States. hal-01193794

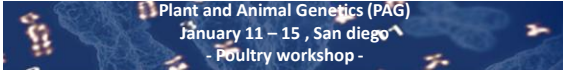
HAL Id: hal-01193794

<https://hal.science/hal-01193794>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Combining Genetics Mapping, Selective Sweeps, Genome Re-Sequencing and Allele Specific Expression Reveals Several Serious Causative Genes for QTL Regions

Pierre-François Roux, Simon Boitard, Anis Djari, Diane Esquerré, Colette Désert, Morgane Boutin, Sylvain Marthey, Frédéric Lecerf, Elisabeth Le Bihan-Duval, Christophe Klopp, Bertrand Servin, Frédéric Pitel, Michel Jean Duclos, Marco Moroldo, Olivier Demeure, Sandrine Lagarrigue



INRA, UMR PEGASE, Rennes, France
Animal genetics department
Agrocampus Ouest, UMR PEGASE, Rennes, France



Context and aim of the study

• **Genetics context**

- During the past 3 decades, thousands of QTL have been mapped by linkage analysis.
- Size of QTL are usually very large, containing hundreds of genes
→ it is quite impossible to identify causative genes & mutations

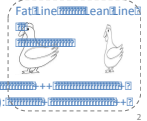
• **Aim**

- Propose a strategy based on detection of selective sweeps in divergent lines (generally used in linkage mapping) to facilitate the identification of causative genes in QTL regions

• **Divergent lines used**

- Fat and lean lines, divergently selected during 7 generations on abdominal fatness (AF) (Lebihan-Duval & Baeza 2013)

- 4 QTL for AF
- 1 QTL for breast muscle Weight (Lagarrigue et al 2006)



1 – Detection of selective sweeps in both divergent lines

• **Re-sequencing data**

- We re-sequenced in 20 X the genome of 11 fat and lean birds + 9 F1 birds used for QTL mapping studies
=> a total of 9.4 M SNP with in average 2.7 M (± 0.5) per bird

• **For detecting signatures of selection**

Genetics, 2013 Mar;193(3):929-41. doi: 10.1534/genetics.112.147231. Epub 2013 Jan 10.
Detecting signatures of selection through haplotype differentiation among hierarchically structured populations.
Fariello M, Boitard S, Naya H, SanCristobal M, Servin B.
Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, Toulouse, France. mfariello@toulouse.inra.fr

- Originality of this approach called **HapFLK** is to focus on the differences of **haplotype frequencies** between populations
And not of **allele frequencies** as usually done in classical studies
- This approach is particularly adapted to our data composed of millions of SNPs

1 – Detection of selective sweeps in both divergent lines - Results -

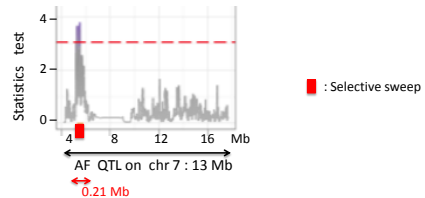
Chr	Chr size (Mb)	Selective sweeps (n)
1	200.99	36
2	154.87	25
3	113.66	17
4	94.23	21
5	62.24	1
6	37.40	5
7	38.38	3
8	30.67	1
10	22.56	6
11	21.93	5
12	20.54	1
13	18.91	1
14	15.82	4
17	11.18	1
20	13.99	1
24	6.40	1
TOTAL	1050.9	129

- ⇒ 129 sweeps on different chromosomes And representing 1.4 % of the genome
- ⇒ The average size of one sweep: 98 kb (± 90)
- Each Sweep contained in average :
⇒ 850 (± 700) SNP
⇒ 2.2 (± 1.5) genes

Do the selective sweeps co-localize with QTL regions ?

Do the selective sweeps co-localize with QTL regions ?

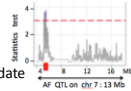
- We overlaid the 129 sweeps with the 5 QTL regions (4 for AF, 1 for BMW)
⇒ All these QTL have at least one selective sweep



- ⇒ Sweep locations reduce drastically QTL sizes (by a factor 50), Reducing drastically the number of candidate genes.

Sweeps reduced drastically QTL sizes

- Among the 5 QTL, 2 QTL correspond to the ideal situation:
 - with only one sweep identified within the QTL
 - containing only one gene
 ⇒ which makes this kind of gene a very serious causative candidate



QTL	BMW-chr1	AF-chr5
QTL size	18 Mb	12
Sweep size	140 kb	100 kb
gene	SGCG	Jag2
Evidence in the literature	yes	No

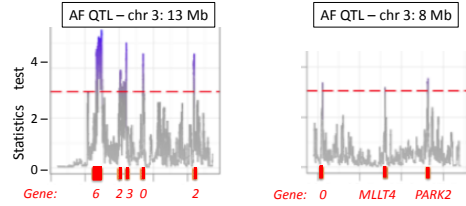
Gene codes for a protein (Sarcoglycan gamma) for which a defect leads to a muscular dystrophy

no evidence was reported about a relationship of this gene and adiposity or lipid metabolism

7

Do the selective sweeps facilitate the identification of causative genes in QTL regions?

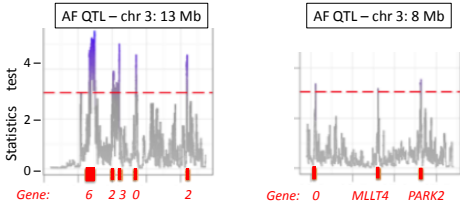
- The situation can be more complicated as shown by the other QTL with several sweeps per QTL and sometimes more than one gene per sweep



8

Do the selective sweeps facilitate the identification of causative genes in QTL regions?

- The situation can be more complicated as shown by the other QTL with several sweeps per QTL and sometimes more than one gene per sweep



9

Can we go further by using different approaches?

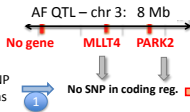
Let's take the example of the AF QTL on chr 3 (8 Mb)
 No gene MLLT4 PARK2

- First, we analyzed the SNP in coding regions of these two genes to detect potential causative mutations:
 - Non synonymous mutation (Variant Effect predictor from Ensembl)
 - highly conserved across species (PhastCons software)
 - Impacting a functional domain (PhastCons software)

For these complex QTL, we propose to go further To highlight the best (one of the best) causative candidate gene...

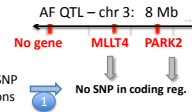
10

Can we go further by using different approaches?

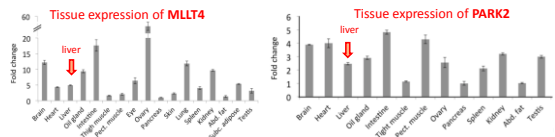


- We analyzed the SNP in the coding regions
 - No SNP in coding reg. ⇒ Suggesting for one of these 2 genes that a causative mutation acts on regulation of its expression

Can we go further by using different approaches?

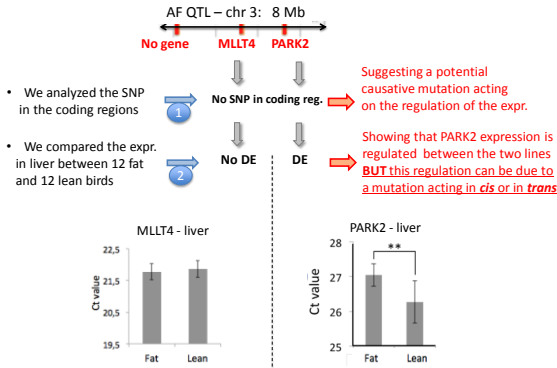


- We analyzed the SNP in the coding regions
 - No SNP in coding reg. ⇒ Suggesting for one of these 2 genes that a causative mutation acts on regulation of its expression
- After checking that these two genes are expressed in liver, a key organ for lipid metabolism
 - We then analysed if they are differentially expressed, in liver, between FL and LL



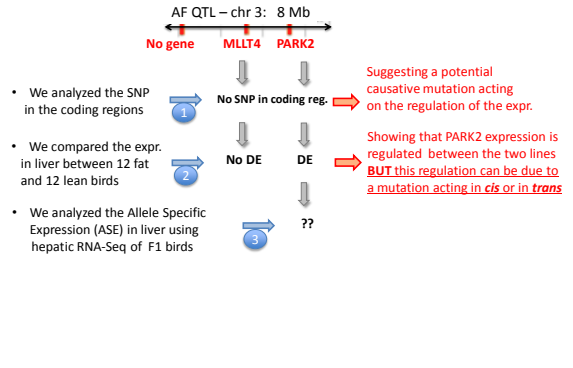
11

Can we go further by using different approaches?



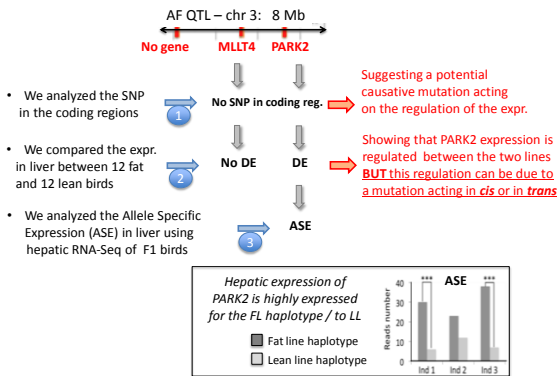
13

Can we go further by using different approaches?

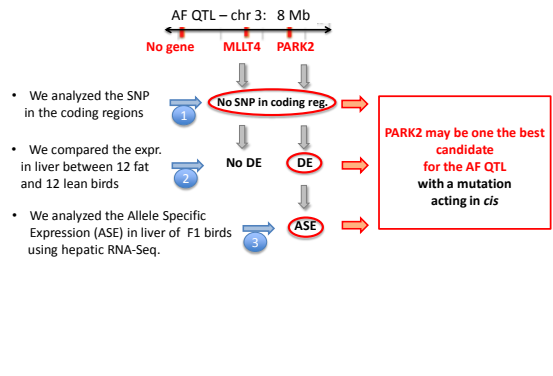


14

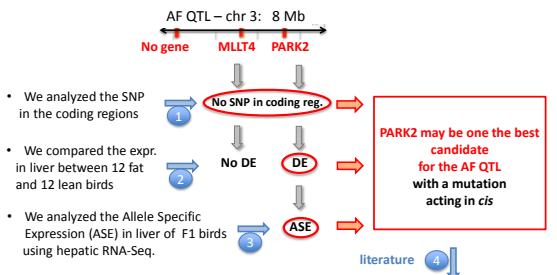
Can we go further by using different approaches?



Can we go further by using different approaches?



Can we go further by using different approaches?



If we compare our results with the literature ...
 PARK2 was until recently only associated to a juvenile form of Parkinson disease.
 More recently (2011) Sack's laboratory showed that PARK2^{-/-} mice are resistant to weight gain induced by high fat and cholesterol diet (HFD) suggesting a role of this gene in adiposity
 => These results in mouse and chicken strengthen the idea that PARK2 is a regulator of fat metabolism

17

In conclusion (first part)

- By combining different approaches as :
 - 1) Detection of selective sweeps and analysis of gene located within (DNA-Seq)
 - 2) Detection of non synonymous SNP in coding regions (DNA-Seq and VeP tool) allowing to focus our research on coding regions versus regulatory regions and then differential expression
 - 3a) Analysis of differential gene expression (DE) in a appropriated tissue (RT-qPCR)
 - 3b) Analysis of allele specific expression (ASE) in a same tissue (RNAseq)
- We identified three serious causative genes underlying BMW or AF QTL
- We identified new function related to lipid metabolism for two of these genes, function never reported so far (IAG2) or not still clearly reported (PARK2)

Limitation:
 The two DE and ASE approaches need to work with appropriated physiological conditions and tissues in which the gene analysed has to be expressed and to impact the trait of interest !!

18

How to go further in exploration of the 129 selective sweeps by using RNA-Seq?

3 aims

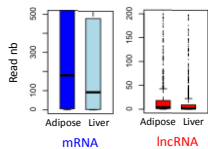
- I. Better explore the **non synonymous SNPs** in coding regions (0.4% of the SNP observed)
- II. Look for systematically **mutations acting in cis** by ASE (using liver and adipose F1 birds) jointly to the analysis of DE of these genes found under ASE (using microarray data related to liver and adipose)
- III. Improve the **genome annotation related to liver and adipose** by detecting new coding genes / transcripts & long non coding RNA (that represents potential regulatory elts for coding mRNA)

→ Some results

19

Some results: Inc RNA in liver and adipose tissue in chicken

- Using a pipeline with different filters :
 - retaining lncRNA with at least 2 exons, allowing us to orientate them
 - no matching with a known exon
 - no having a predicted coding structure
- => we found 1,750 lncRNA among them 66% have at least 3 reads in liver and/or adipose



As already reported in human (Derrien et al 2012) the expression level of lncRNA is very low compared to the coding mRNA

Some results: New coding genes/transcripts in liver or adipose

- We found 10,486 new transcripts (corresponding to 6,587 known genes) expressed in liver or adipose (→ quite a lot compared to the 17000 referenced in gal4 annotation)
- We found 4,109 new genes corresponding to 5,289 new transcripts with 5 % and 15% specific to liver and adipose respectively

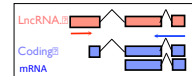
=> We are currently overlying these new genes with the 37 selective sweeps without gene

20

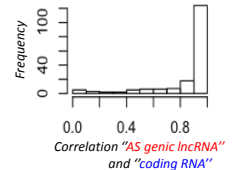
Some results: Inc RNA in liver and adipose tissue in chicken

- Most of the lncRNA (83%) are in intergenic regions as already reported in human (Derrien et al 2012)

- Among the genic lncRNA, most of them share at least one exon of coding mRNA associated, in AS orientation

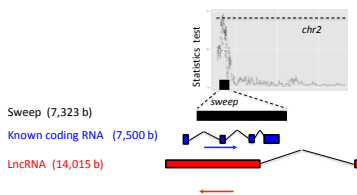


Most of these AS genic lncRNA are highly positively correlated with the coding RNA associated (using the 28 samples available)

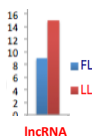


22

We overlaid these lncRNA with the 82 sweeps with no or one gene



- These two coding RNA and lncRNA are expressed in adipose and liver
 - Their expression are highly correlated across the 28 samples ($r=0.97$)
 - ncRNA tends to be less expressed in liver of FL compared to LL that could impact the translation of coding RNA and affect the trait
- => We are currently analyzing if this lncRNA is regulated in cis by analyzing its ASE using liver of F1 birds



23

Sharing , sharing, sharing ... with the chicken community

- Because the lncRNA are very lowly expressed and most of them are shared by # tissues it is interesting to accumulate a lot of RNAseq data to have high depth in order to better observe them

=> Our RNAseq data from liver and adipose have been shared with data from other tissues within the avian RNA-Seq consortium framework driven by the Roslin institute in order to have a better view of the global genome annotation in chicken. (Jacqueline Smith and Dave Burt as co-ordinators)

=> One first step towards the AgEncode project ...

- We can also share the genome sequences to improve the genome assembly and variant catalog... so our 20 genome sequences in 20X (HiSeq2000) are also available

24

Many thanks

⇒ INRA, LGC, Toulouse, France Simon Boitard (HapFLK) Bertrand Servin Frédérique Pitel	⇒ INRA, UR 38, Avian research, Tours, France Elisabeth Le Bihan – Duval (for the FL and LL lines)
⇒ INRA, SIGENAE, Toulouse, France Anis Djari Christophe Klopp	⇒ INRA, UMR GABI, CRB GADIE, Jouy-en-Josas, France Sylvain Marthey Marco Moroldo Jordi Estelle (DNA-Seq obtained by capture)
⇒ INRA, Genotoul Plate-form, Toulouse, France Diane Esquerre (DNA-Seq and RNA-Seq generation and treatment)	⇒ CNRS, IGDR, Rennes, France Thomas Derrien Christophe Hitte (long ncRNA annotation)
⇒ INRA - Agrocampus Ouest, - UMR PEGASE, Rennes, France Genetics & Genomic Team Pierre francois Roux - PhD Colette Désert Frederic Lecerf Olivier Demeure Morgane Boutin Sandrine Lagarrigue	

25

Many thanks

INRA – Agrocampus University - Rennes, France Genetics & Genomic Team Pierre francois Roux - PhD Colette Désert Frederic Lecerf Olivier Demeure Morgane Boutin Sandrine Lagarrigue		
--	--	--

26

Annexes

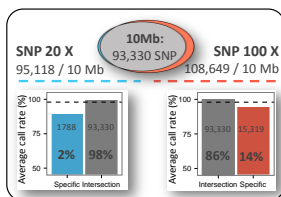
1 - Detect SNP and selective sweeps in F0 fat and lean lines

- Re-sequencing (HiSeq2000) the genome of 11 F0 fat and lean birds + 9 F1 birds used for QTL mapping studies → sequencing depth of 20 X
- Developmnt of a robust procedure for identifying reliable SNP (steps)
 - Alignment on WASHUC2.1 BWA
 - Filter on mapping quality Samtools
 - Removal of PCR duplicates Samtools
 - Realignment / Recalibration & SNP calling GATK
- Results
 - 9.4 M reliable SNP
 - with 2.7 M (± 0.5) SNP per individual (2.6 SNP/ kb)
- Annotation of these SNP with Variant effect Predictor package (Ensembl):
 - 90.5% = intergenic (48.7 %) or Intronic (41.8 %) regions
 - 8.3% = 2kb upstream or downstream of genes
 - 1.2% = Exonic regions among them:
 - 25,7 % lead to a misense effect or a modification in a start or stop codon

27

Is sequencing depth of 20X sufficient to observe the whole SNP present in each individual ?

Eleven individuals have been sequenced on a 10 Mb window
 With 20 X depth (obtained in present data)
 and 100 X depth (obtained by DNA capture)



⇒ 20 X : 98 % of SNP were in intersection
 2 % specific
 But specific SNP are, in average, of poor quality (average call rate < 90 %) and must be considered as false positives

⇒ 100 X : 86 % of SNP were in intersection
 14 % specific
 Specific SNP have an average call rate of 92 % i.e. among them there are false positives and reliable SNP

In ccl : 20 X re-sequencing is sufficient to have information for almost all SNP

29

HapFLK (Fariello et al 2013)

- 1) - Definition of haplotypes locally by clustering individuals mixing the populations (FastPHASE – Sheet and tephens, 2006) => haplotype cluster
 - Calculation of haplotype frequencies in each population
- 2) - For each Haplotype, calculation of a measure of “genetic differentiation inter population” (~Fst)
- 3) - Estimation of the distribution of this measure across all genome (neutral haplotype)
 - Identification of haplotypes which are extreme compared to this distribution (haplotype on selection)

hapFLK

- For each SNP, calculation of a measure of "genetic differentiation inter population"

$$F_{ST} = \frac{s_p^2}{\bar{p}(1-\bar{p})} = \frac{1}{n-1} \frac{\sum_{i=1}^n (p_i - \bar{p})^2}{\bar{p}(1-\bar{p})}$$

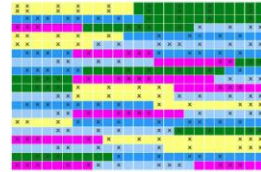
- Estimation of the distribution of this measure for all the neutral SNP
- Identification of SNP which are extreme compared to this distribution

31

Test hapFLK (Fariello et al, 2013)

Version haplotypique de FLK :

- Clustering local des individus à l'aide du logiciel FastPHASE (Sheet and Stephens, 2006), en mélangeant toutes les populations.
- Pour chaque SNP, calcul des fréquences de cluster par population et utilisation d'une version multi-allélique de FLK.



32

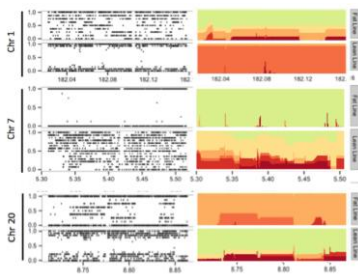
1 - Detect SNP and selective sweeps in F0 fat and lean lines

Genetics, 2013 Mar;193(3):929-41. doi: 10.1534/genetics.112.147231. Epub 2013 Jan 10.

Detecting signatures of selection through haplotype differentiation among hierarchically structured populations.

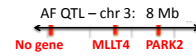
Fariello M, Boillard S, Naya H, SanCristobal M, Servin B.

Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, Toulouse, France. mfariello@toulouse.inra.fr



33

Can we go further by using different high throughput data?



- We analyzed the SNP in the coding regions → No SNP in coding reg. → Suggesting a potential causative mutation acting on the regulation of the expr.
- We compared the expr. in liver between 12 fat and 12 lean birds → No DE / DE → Showing that PARK2 expression is regulated BUT this regulation can be due to a mutation acting in cis or in trans

