

INRAE

RNA-seq data for detecting reliable SNPs & genotypes in livestock species: interest for coding variant characterization and cis-regulation analysis by allele-specific expression.

Jehl et al, in Frontiers In Genetics 2021

+ updating in 2022 C. Guyomar² M Charles et al²

F Jehl^{1,2†}, F Degalez^{1,2†}, M Bernard^{2†}, F Lecerf^{1,2}, L Lagoutte^{1,2}, C Désert^{1,2}, M Coulée^{1,2}, O Bouchez³, S Leroux⁴, B Abasht⁵, M Tixier-Boichard⁶, B Bed'hom⁶, T Burlot⁷, D Gourichon⁸, H Acloque⁶, S Foissac⁴, S Djebali⁴, E Giuffra⁶, T Zerjal⁶, F Pitel⁴, C Klopp² & Sandrine Lagarrigue^{1,2}

^{2, 3, 4, 6}INRAE, ¹INSTITUT AGRO, ⁷NOVOGEN, France & ⁵Univ. of Delaware

Context



- For detecting polymorphisms in the whole genome of a population, **DNA-seq data** analyzed by the **bioinformatics GATK tool is the standard approach.**




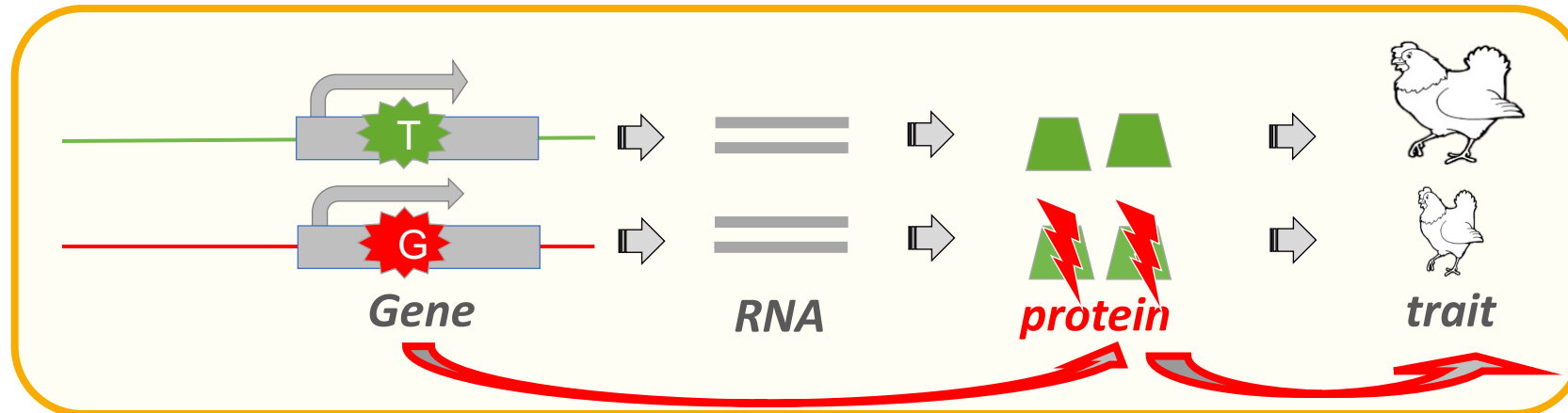
- DNA-seq data are expensive to generate and to store, because these data are large

Context

- **RNA-seq data**, are cheaper to generate & are generally used for studying gene expression,
→ can also be used to detect genomic variations but in **only expressed regions of the genome**,

It can be interesting to work on these specific regions

- 
- 1) **For characterizing variants** affecting the structure/function of associated proteins which might be a causative variants of traits or disease



Context

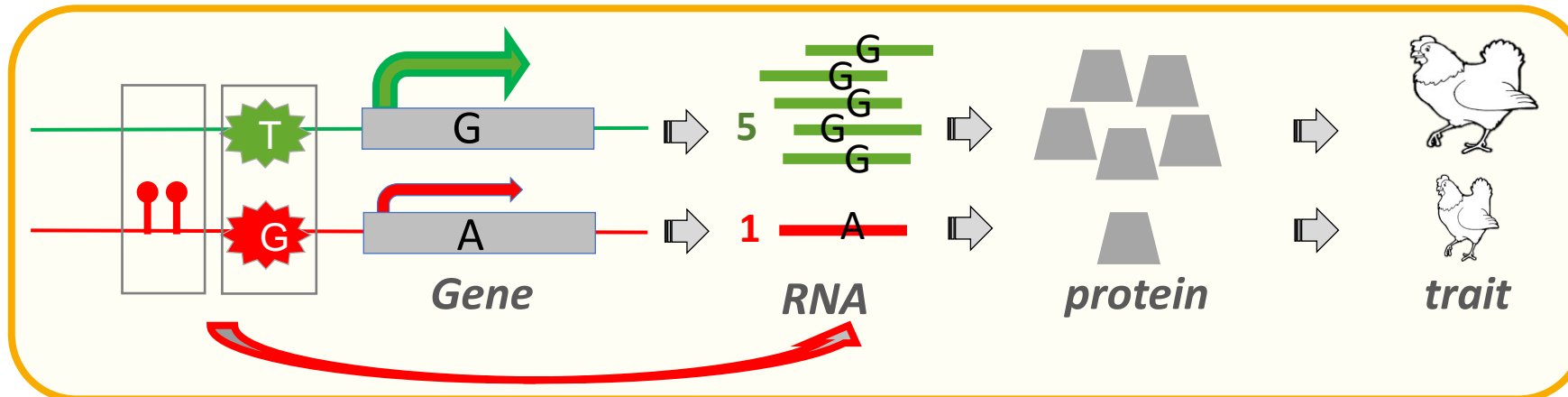
- **RNA-seq data**, are cheaper to generate & are generally used for studying gene expression,
→ can also be used to detect genomic variations but in **only expressed regions of the genome**,

It can be interesting to work on these specific regions



1) **For characterizing variants** affecting the structure/function of associated proteins
which might be a causative variants of traits or disease

2) **For characterizing a variation of RNA level between the two chromosomes of animal**
These RNA level variations can be due to an epigenetic variation or a genetic regulatory variant (cis-eQTL)



Context

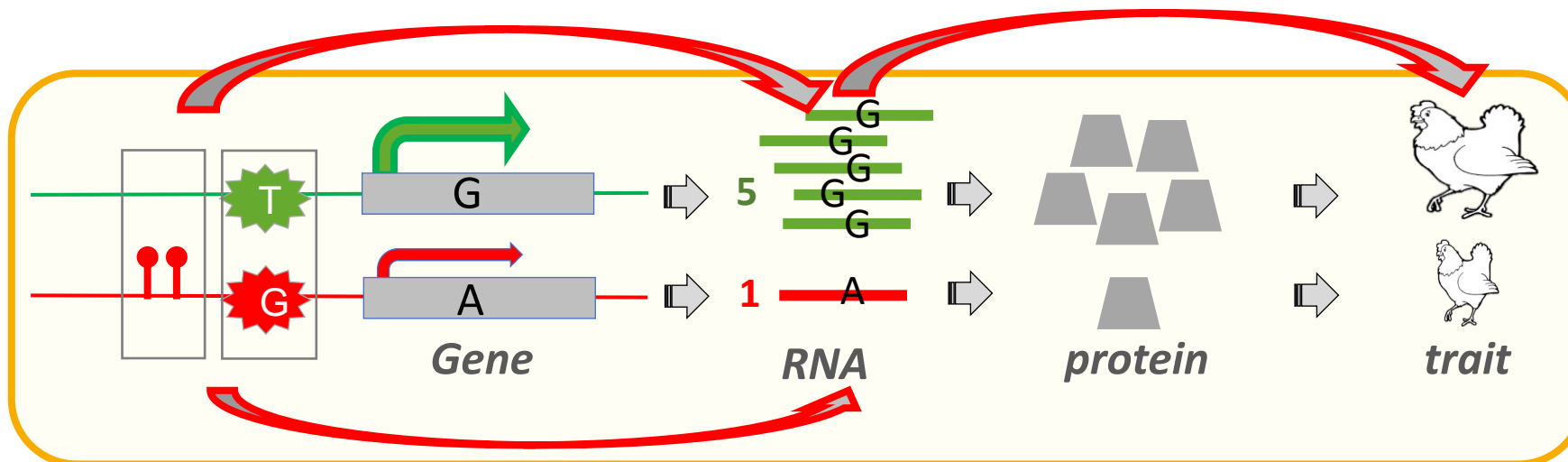
- **RNA-seq data**, are cheaper to generate & are generally used for studying gene expression, → can also be used to detect genomic variations but in **only expressed regions of the genome**,

It can be interesting to work on these specific regions

- 1) **For characterizing variants** affecting the structure/function of associated proteins which might be a causative variants of traits or disease
- 2) **For characterizing a variation of RNA level between the two chromosomes of an animal**

The analysis of **allele-specific expression (ASE)** can be useful :

- **for identifying regulatory causative variants of trait or disease** (Orozco et al, 2020, Le bihan-Duval et al, 2011)



Context

- **RNA-seq data**, are cheaper to generate & are generally used for studying gene expression, → can also be used to detect genomic variations but in **only expressed regions of the genome**,

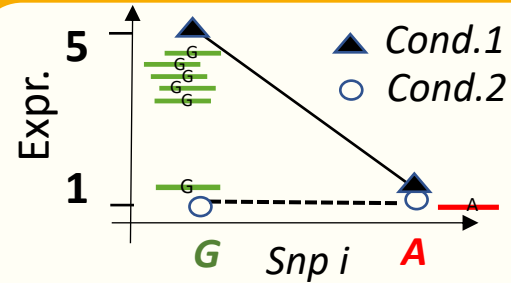
It can be interesting to work on these specific regions



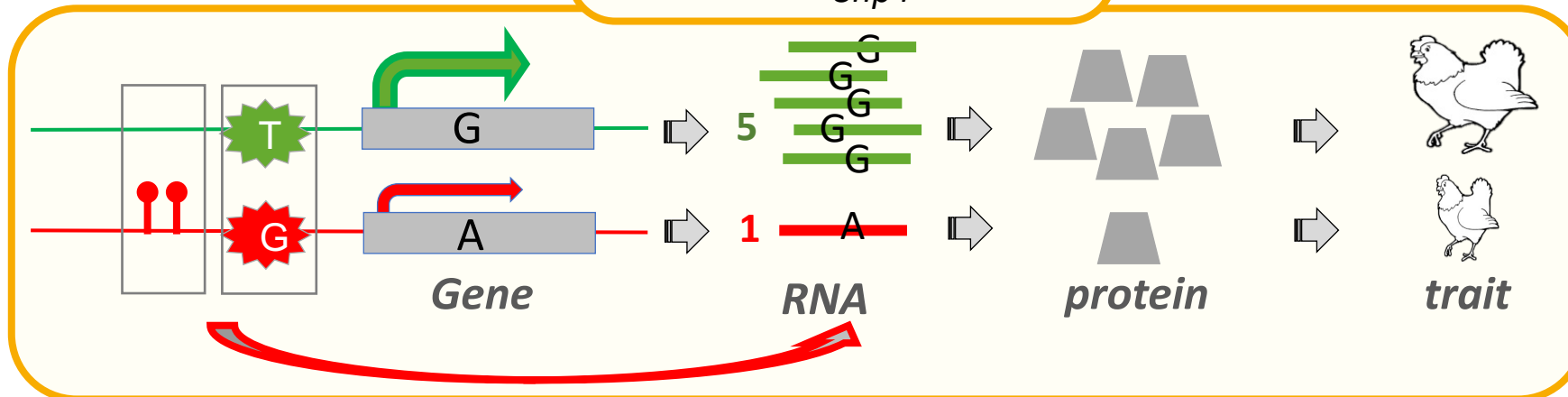
- 1) **For characterizing variants** affecting the structure/function of associated proteins which might be a causative variants of traits or disease
- 2) **For characterizing a variation of RNA level between the two chromosomes of an animal**

The analysis of **allele-specific expression (ASE)** can be useful :

- for studying interaction **G x E** as the **ASE** can be condition-specific



Moyerbrailean et al, 2016
Kim-Hellmuth et al. 2017,



Context

- **RNA-seq data**, are cheaper to generate & are generally used for studying gene expression,
→ can also be used to detect genomic variations but in **only expressed regions of the genome**,

It can be interesting to work on these specific regions



- 1) **For characterizing variants** affecting the structure/function of associated proteins
which might be a causative variants of traits or disease
- 2) **For characterizing a variation of RNA level between the two chromosomes of an animal**
The analysis of **allele-specific expression (ASE)** can be useful :



- A huge amount of RNA-seq data have been accumulated over the last 10 years
→ Allowing to analyze genomic variant in different populations of various species
- In general, several dozen of animals are sequenced per population
→ allowing to estimate frequencies of genotypes and alleles

Context

- **Gene expression can be variable :**

- between genomic positions, **position 3 is not sequenced enough to reveal properly a SNP**

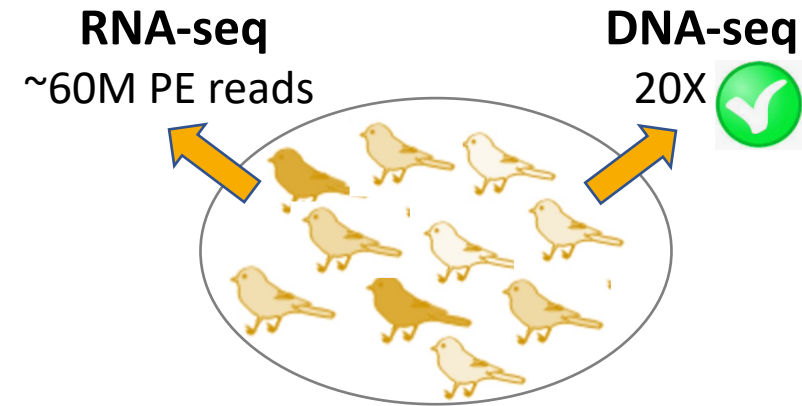
- between individuals, depending on the conditions in which animals are
(for the SNP in position 4, there is a very low expression in stressful condition compared to the control)
 → can lead to **missing genotypes or false positives** because not supported by enough observations

1 tissue

position		Pop.	Observation number						Genotype (GT)						
			ind.1	ind.2	ind.3	ind.4	ind.5	ind.6	ind.1	ind.2	ind.3	ind.4	ind.5	ind.6	
1:	SNP1 (gene1)	✓	>20	100	60	80	50	15	35	0/0	1/1	0/0	0/1	1/1	0/0
2:	SNP2 (gene1)	✓	>20	1200	1500	880	1700	1500	980	0/0	0/1	1/1	0/1	1/1	0/0
3:	reliable SNP?	✗	10	2	3	0	1	4	0	./.	./.	./.	./.	.1/1?	./.
4:	SNP3 (gene3)	✓	>20	14	15	13	1	2	4	0/1	0/1	1/1	./.	./.	0/0?
				Control			Stressful cond.			Control			Stressful cond.		

The first Objective (*Jehl et al, 2021, Frontiers in Genetics*)

- To develop a workflow to detect SNPs & Genotypes from RNA-seq data, based on GATK tool & to provide some performances of RNA-seq in terms of precision and sensitivity
- We used 2 independent populations in which RNA-seq and DNA-seq were available for the same birds & we assumed that the DNA-seq data represent the “truth”



- 1) Layer pop. with **n=15**
- 2) Broiler pop. with **n = 8**

The first Objective (Jehl et al, 2021, *Frontiers in Genetics*)

- To develop a workflow to detect SNPs & Genotypes from RNA-seq data, based on GATK tool & to provide some performances of RNA-seq in terms of precision and sensitivity
→ We used 2 independent populations in which RNA-seq and DNA-seq were available for the same birds & we assumed that the DNA-seq data represent the “truth”

- As a main result,
we show for SNP detection by RNAseq :

- a precision superior to 90%
- a recall (sensitivity) which depends on the individual & tissue number which are analysed

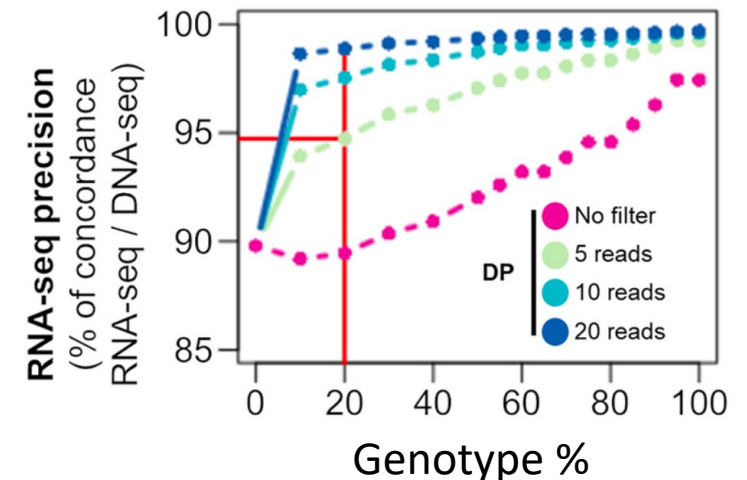
With 15 birds and 1 tissue analyzed, the recall=85% (SNP detected by DNAseq in expressed regions were detected by RNAseq)

- For the genotype detection by RNAseq ;

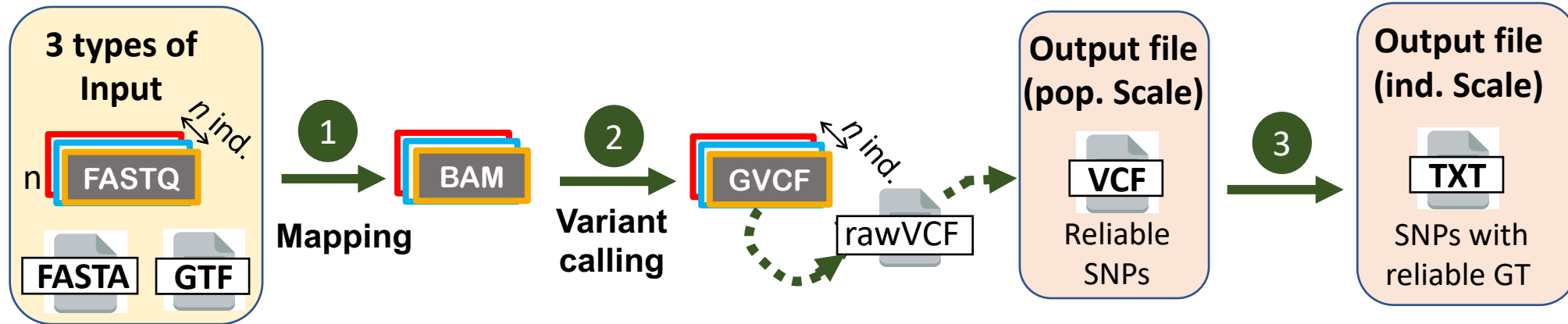
- a precision of 90% without using filters on the reads supporting the genotype

→ 95% with at least 20% of genotypes supported by ≥ 5 reads

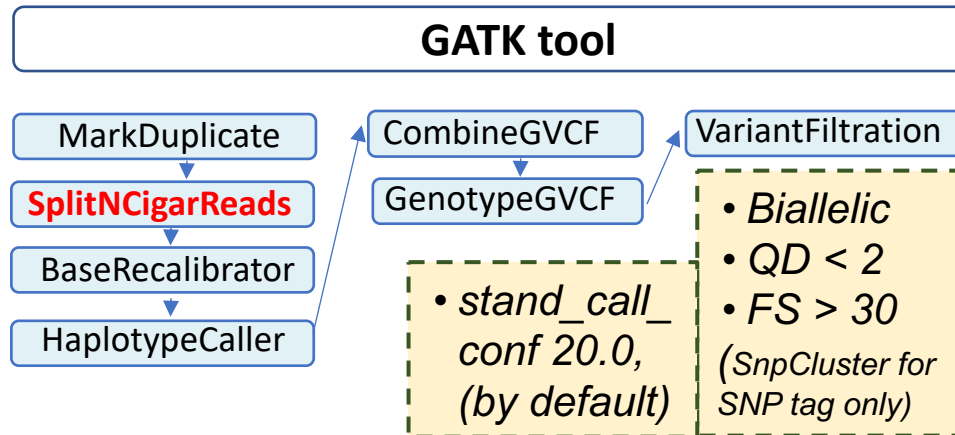
→ over 97% when supported by ≥ 10 reads



Workflow for detecting SNPs & Genotypes (GT)



STAR
2-pass



Home-made code

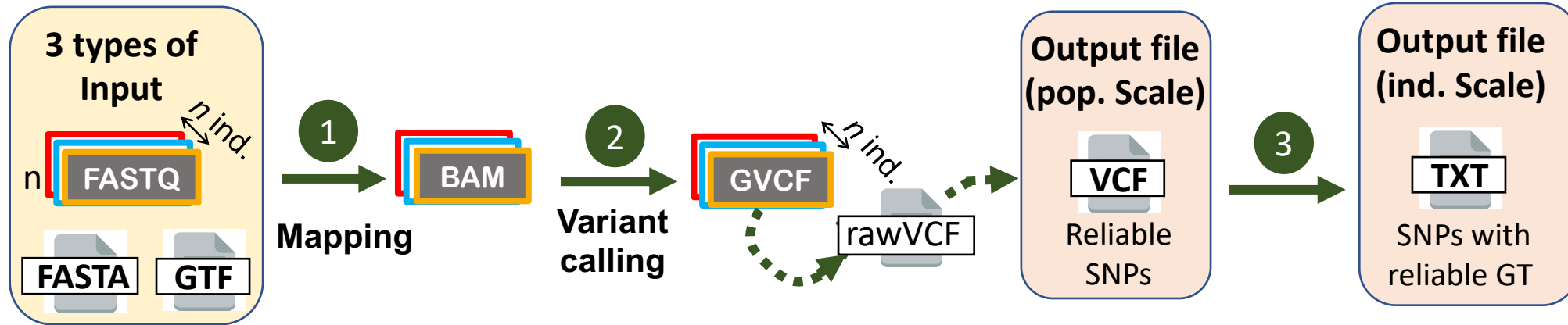
- *Read nb / genotype*
- *Genotype nb / SNP position*

These filters can be defined by the user who has to do a compromise :
'The more reliable GT you have per SNP, the less SNPs you have'

FASTA = reference chicken genome

GTF = genome annotation (providing gene and transcript positions)

Workflow for detecting SNPs & Genotypes (GT)



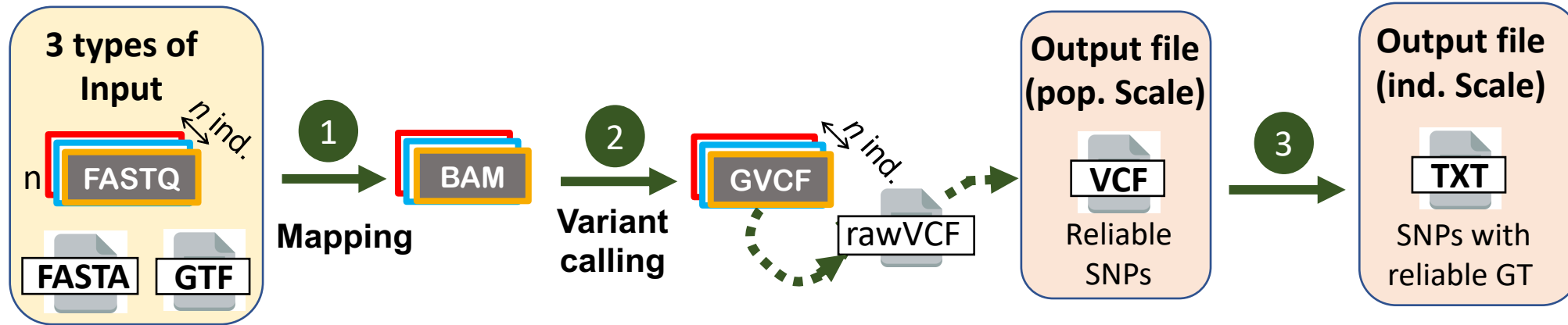
Applying to 10 available chicken pop. (380 indiv. ~800 RNAseq), we detected:

~10M SNPs, [23 M SNPs in the *Ensembl dbSNP database (v107)*

With on average **~ 1.5 M** per population

Among them **~500,000 SNPs** having reliable genotypes observed in at least **50%** of the population

Workflow for detecting SNPs & Genotypes (GT)



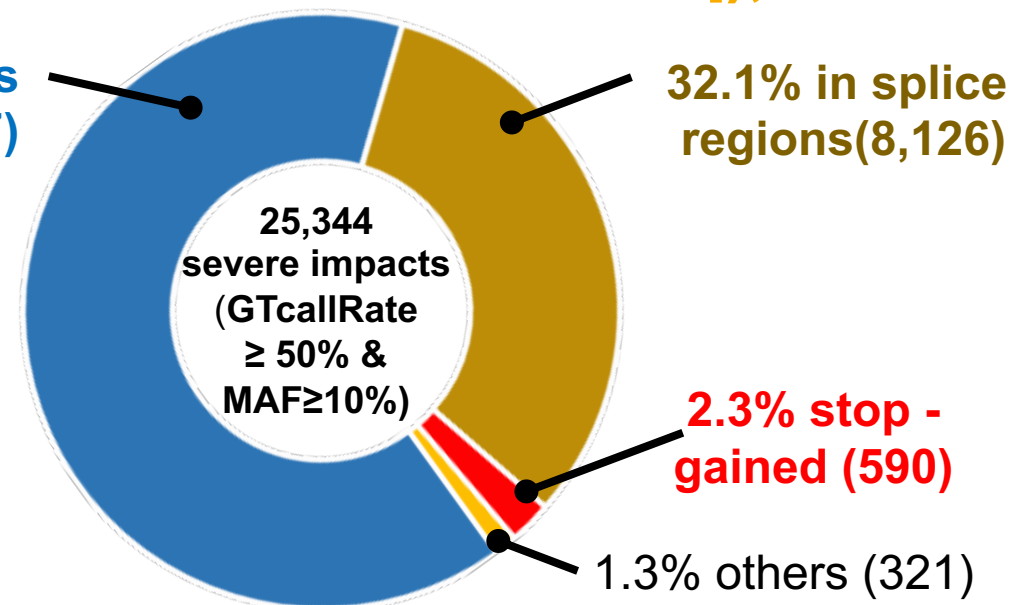
Applying to 10 available chicken pop. (380 indiv. ~800 RNAseq), we detected:

~10M SNPs

~ 25,000 deleterious* variants observed in at least one population

& found for them a lower frequency of the homozygotes DEL/DEL compared to the synonymous variants

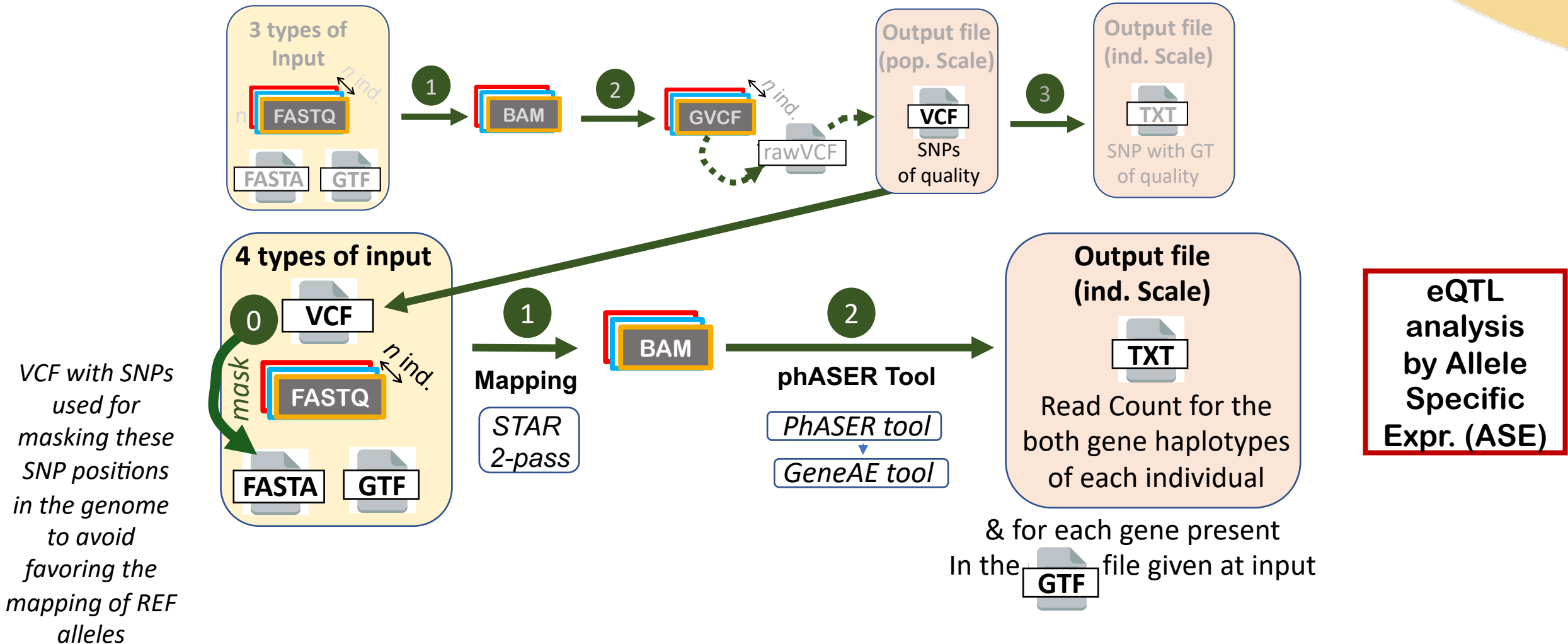
64.3% SIFT-deleterious missenses (16,307)



*with a MAF (here \geq 10%) in at least one population to remove spurious deleterious allele due to sequencing errors

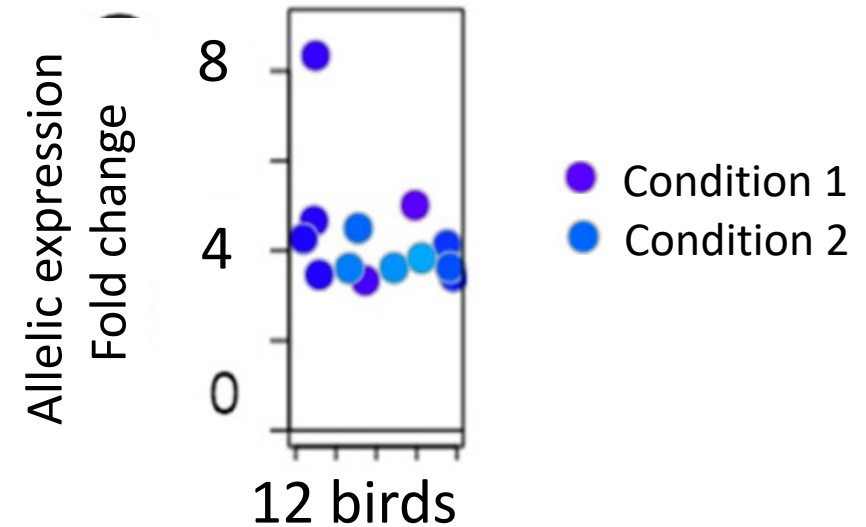
The 2nd objective : Workflow for Allele Specific Expression (ASE) based on the phASER tool (S Castel)

The originality of phASER is to phase heterozygous variants of one gene and then aggregates counts across variants within the gene to produce a single expression measurement for each haplotype



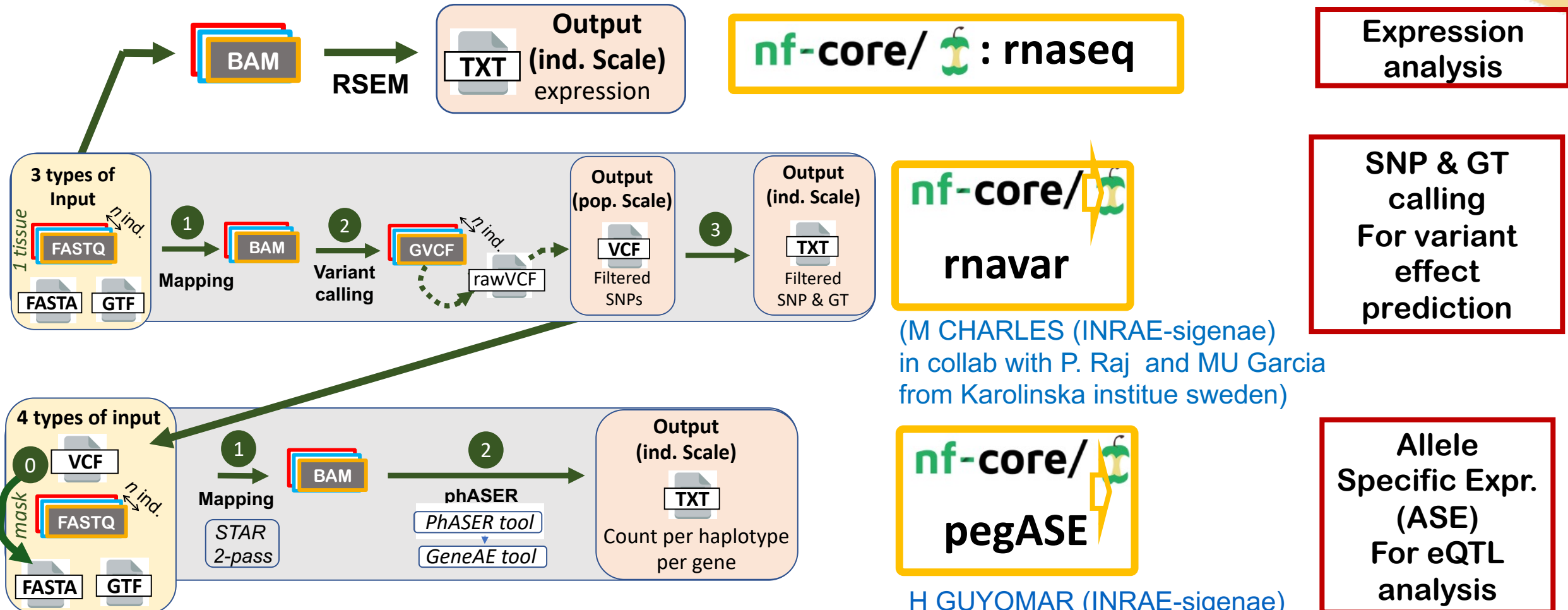
Application of the workflow 2 on two populations using one tissue (liver)

- Among the ~11000 protein-coding genes & ~3000 lncRNA genes expressed in liver (TPM ≥ 1)
→ We observed on average 8-10 SNPs per gene, allowing to analyse Allelic Expression
→ We found ~30% of genes which are cis-regulated
- As illustration, the gene **ACOT1L** has an allelic expression Fold Change (FC) between the two haplotypes of 12 birds of **4 fold**



The next step is to analyze interaction between 'G X E by analysing the expression unbalance according to conditions in the populations, in which we have control and stressful conditions (as heat stress, diet changes, age,...)

These 2 workflows already available in Jehl et al, 2021, Frontiers in genetics, will be soon available in the nfCore platform, in which pipelines are standardized for reproducible analysis

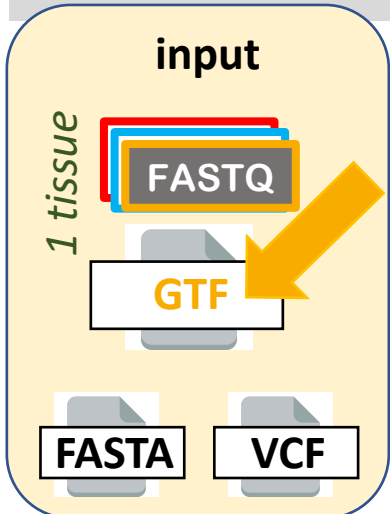


(M CHARLES (INRAE-sigenae) in collab with P. Raj and MU Garcia from Karolinska institute sweden)

H GUYOMAR (INRAE-sigenae) <https://github.com/cguyomar/nf-pegASE>

In the 2 workflows, a GTF genome annotation file is required the more genes, the richer the analysis

Workflow 1 Workflow 2



- A new chicken genome reference, **GRCg7b**, came out in June with a new associated **Ensembl gene annotation v107**

<i>GRCg7b</i>	e! v107	→	e! v107 'extended'	<i>Degalez et al, 2022, in preparation</i>
<i>chicken</i>	17,007	x1.5	25,137 Protein Coding Genes	
<i>genome</i>	11,944	x3.8	45,699 Long non coding Genes	
	~29,000		~71,000 genes	

→ we have enriched this annotation v107, by adding new genes using other resources like REFseq (NCBI), NONCODE, FAANG resources

→ **Increasing the gene number by 4 times**

→ This new atlas is available at <http://www.fragencode.org/>



In which there are previous Ensembl annotations that we enriched with the same approach

<i>GRCg6a</i>	e! v101	→	e! v101 'extended'	<i>Jehl et al, 2020, scientific reports</i>
<i>Galgal5</i>	e! v94	→	e! v94 'extended'	

Data contributors

Workflow + analysis

INRAE SIGENAE TEAM

M. BERNARD
M. CHARLES
C. GUYOMAR
C. KLOPP

INRAE & INSTITUT AGRO GENETICS & GENOMICS TEAM

F. DEGALEZ
F. JEHL
F. LECERF
K. MURET
S. LAGARRIGUE

INRAE & INSTITUT AGRO FR-AGENCODE

H. ACLOQUE
S. DJEBALI
S. LAGARRIGUE
E. GIUFFRA
S. FOISSAC

L. LAGOUTTE (INRAE)
C DESERT (INSTITUT AGRO)
O BOUCHEZ (INRAE)
S LEROUX (INRAE)
B ABASHT (Univ DELAWARE – Newark)
M TIXIER-BOICHARD (INRAE)
B BED'HOM (INRAE)
T BURLOT (NOVOGEN)
D GOURICHON (INRAE)
T ZERJAL (INRAE)
F PITEL (INRAE)



Download the new chicken annotation

- This new atlas is available at <http://www.fragencode.org/>
- Jehl et al, *Frontiers in genetics* 2021
- *Rnavar*
- *pegASE* (<https://github.com/cguyomar/nf-pegASE>)



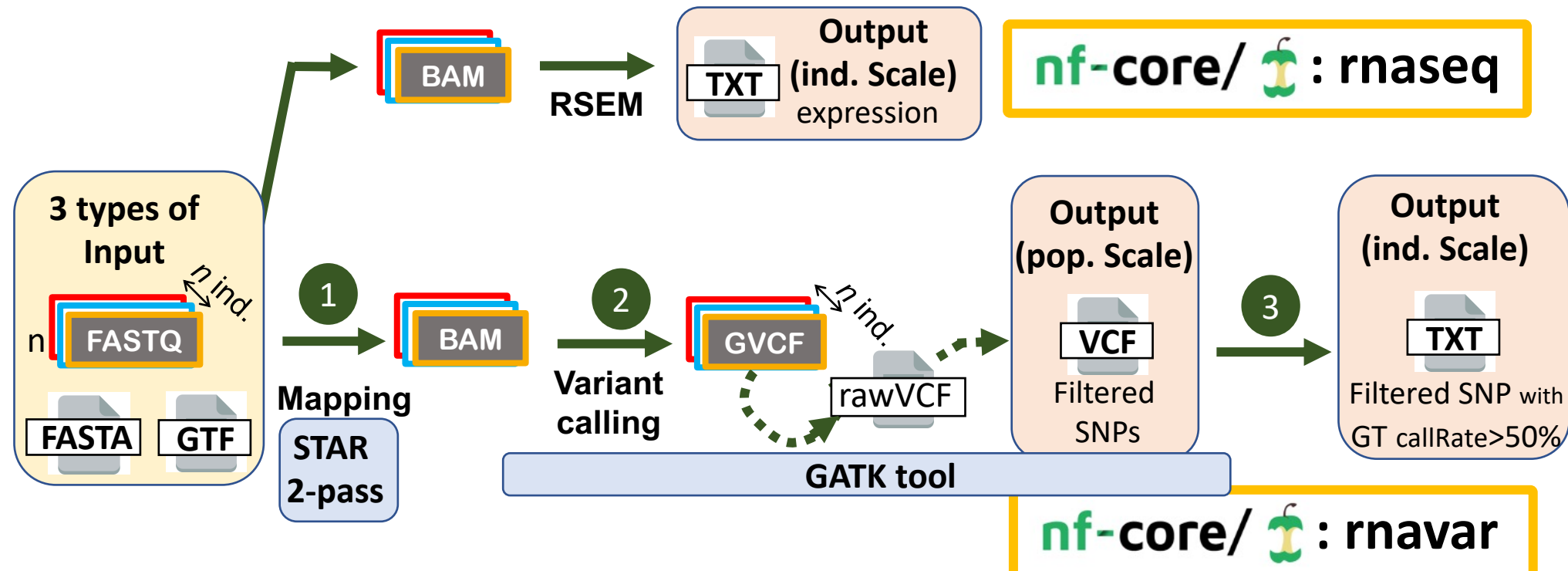
Project fatInteger
Project ChickStress
Project EFFICACE



Horizon2020
Project Feed-a-Gene

ANNEXES

Workflow for Allele Specific Expression (ASE) (Jehl et al 2021)



- This workflow coded in Snakemake language will be soon available in nextFlow language in the nfCore platform in which the workflow for RNAseq expression is already available

→ We are collaborating with P. Raj and M. U. Garcia who have developed such a nfCore workflow 'rnavar' Our aim is to improve some caveats of this first version 1.0 :

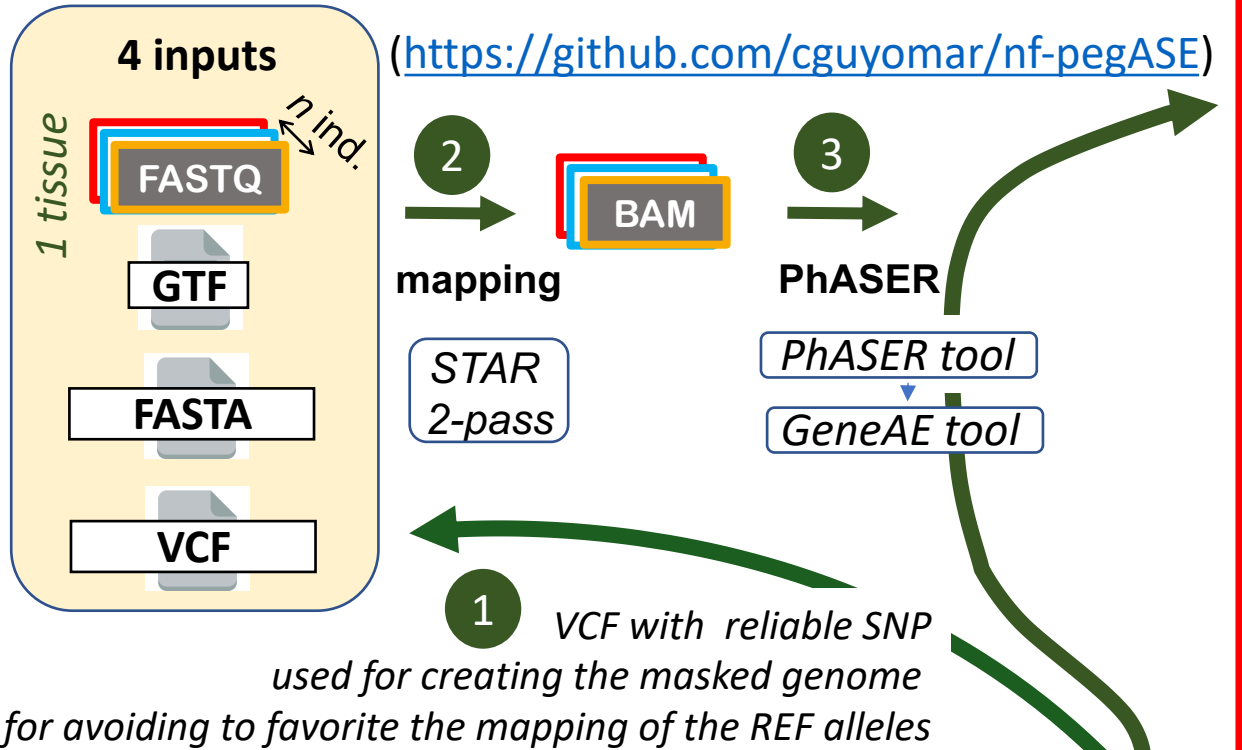
- i) The tool do not generate GVCF, preventing to agregate new samples into a VCF.
- ii) Temporary files management is not very efficient (to run one FASTQ sample, we need 5-8x fastq size
- lii) Rnavar does not perform 2pass mapping, which improves alignment of spliced reads

Workflow for Allele Specific Expression (ASE)

Workflow 2 'ASE'

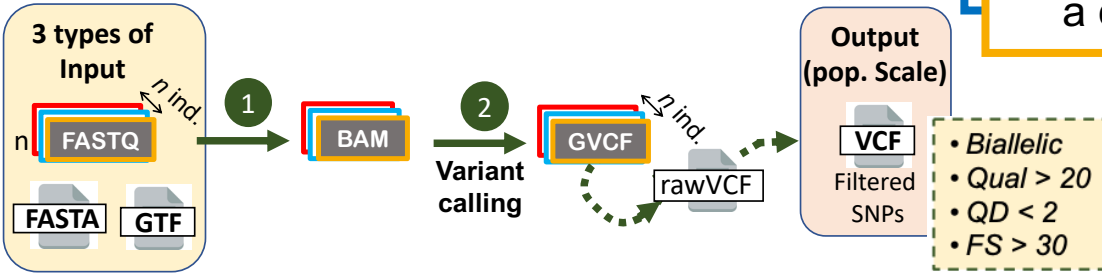
- We developed a workflow based on the phASER tool (Stephen Castel)
- [nf-core/pegase](https://nf-core.org/pegase) A nfCore version, 'nf-pegASE' is in development

(<https://github.com/cguyomar/nf-pegASE>)



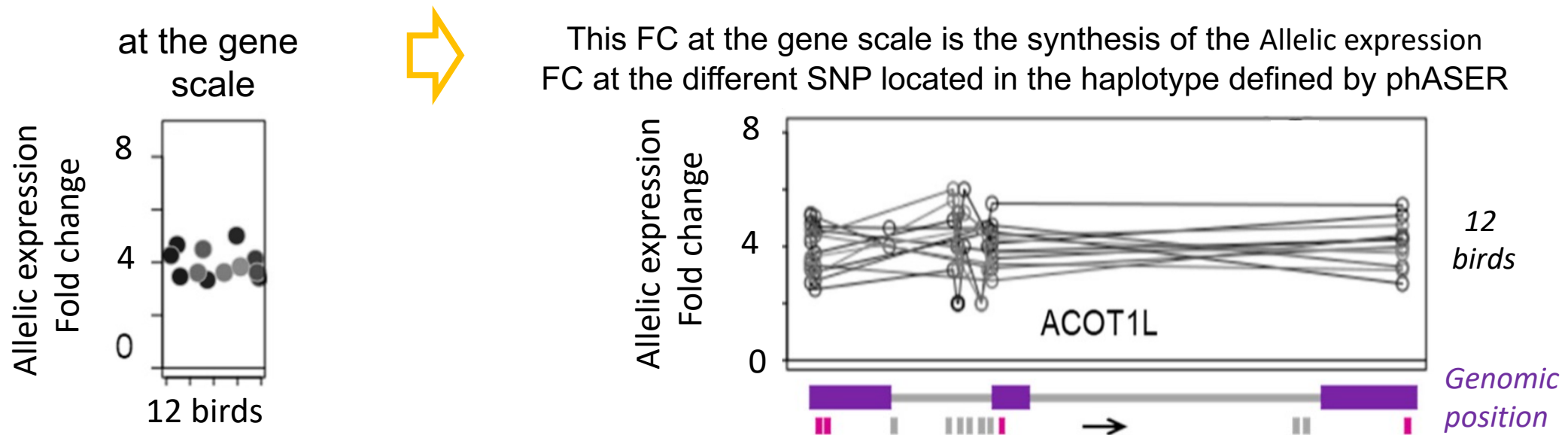
- In each sample of the tissue of interest, phASER **phases SNPs** from the user-provided VCF, using the reads from the BAM file. => haplotype-Count
- PhASER leads to a **list of SNP combinations** upon which phASER **counts** the nb of reads associated to each "super-allele" (haplotype) of these SNP combinations.
- PhASER Gene AE **selects one SNP combination** per gene (i.e. having the haplotype with the most of reads), using the GTF file.
→ we considered only the genes represented by a SNP combination with a haplotype with at least 10 reads.
- Finally, the read number imbalance (ASE) between the haplotypes are performed using a binomial test followed by a correction for multiple tests

Workflow 1 'RNAvar'



Application of the workflow 2 on two populations using one tissue (liver)

- Among the ~11000 protein-coding genes & ~3000 lncRNA genes expressed in liver (TPM ≥ 1)
→ We observed on average 8-10 SNPs per gene allowing to analyse Allelic Expression
→ We found ~30% of genes which are cis-regulated
- As illustration, the gene **ACOT1L** has an allelic expression Fold Change (FC) between the two haplotypes of 12 birds of **4 fold**



In Perspective: In order to identify interaction between 'G X E', we will analyse the expression unbalance in the different populations, in which there are control and stressful conditions