



**Identification of functionally active genomic features relevant to phenotypic diversity
and plasticity in cattle**

Deliverable 3.1

Audit on required and available pipelines across the consortium

Grant agreement no°: 815668

Due submission date

2020-02-29

Actual submission date

2020-02-28

Responsible author(s)

FMV; andreiaamaral@fmv.ulisboa.pt

Confidential No

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668. The content of this report reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

DOCUMENT CONTROL SHEET

Deliverable name	Audit on required and available pipelines across the consortium
Deliverable number	D3.1
Partners providing input to this Deliverable	CRG, FMV, LUKE
Draft final version circulated by lead party to: On date	All partners in WP or task
Approved by (on date)	FBN as Coordinator (2020-02-29)
Work package no	
Dissemination level	Public (PU)

REVISION HISTORY

Version number	Version date	Document name	Lead partner
V1	2020-02-21	D3.1	FMV
V2	2020-02-24	D3.1-vs2-FBN	FMV
V3	2020-02-28	D3.1-vs3-FBN	FMV

Changes with respect to the DoA (Description of Action)

As part of the effort of clustering activities of the three FAANG projects (BovReg, AquaFAANG and Gene-Switch), a second version of the deliverable is envisaged as a result of the actions to be decided at the “FAANG Shared Workshop: Foundation for the future agenda “ that occurred 25th to 27th February 2020.

Dissemination and uptake

This is a deliverable which should be in public domain, therefore available for consultation in the BovReg website. Furthermore, sharing with the other FAANG consortiums is desirable as part of the Bioinformatics clustering activities.

Table of Content

1. Summary of results	2
2. Introduction	3
3. Core report.....	4
3.1 Methods	4
3.2 Results	5
3.2.1 Section 2.....	5
3.2.2 Section 3.....	10
3.2.3 Section 4.....	15
3.2.4 Section 6.....	18
3.2.5 Section 7.....	20
3.2.6 Section 8.....	24
3.2.7 Section 9.....	28
3.2.8 Section 10	28
3.2.9 Section 11	33
4. Conclusions.....	37
5. References.....	39
6. Annexes	46

1. Summary of results

The successful achievement of BovReg's aims in the field of bioinformatics depends strongly on the development of reference annotation pipelines, which guarantee the reproducibility and robustness of BovReg data analyses. Furthermore, BovReg partners will need to integrate several analysis steps and different types of data, each providing different information regarding different layers of genome complexity, to enable a better understanding of the architecture and the regulation processes of the Bovine genome.

Therefore, a survey was developed to identify which tools are being currently used by BovReg partners in different workflows as well as in which type of platform and with which type of workflow environments bioinformatic analyses are performed.

This survey as elaborated in this deliverable therefore enables the identification

- 1) of tools that should be included in the development of standardized and normalized pipelines as well as
- 2) of the need to develop novel standardized functions.

A survey organized in 11 sections with 36 questions was developed by WP3 experts involved in WP3, Task3.1, of which not all participants had to fill in all sections. It was developed taking into account the state of the art of bioinformatics tools for annotation

and analyses of next generation sequencing (NGS) data and the practical experience of the BovReg parties.

A total of 13 responses were obtained from BovReg partners, covering all of those involved in the scientific work packages of BovReg as data analysts. Overall, the survey allowed identifying in detail which tools are being used at different stages of analyses within workflows. Furthermore, the survey showed that for some steps, BovReg partners mostly rely on their own developed analysis code, thus suggesting the need for the development of standardized functions for filtering for data quality and for evaluation of complexity of libraries for describing epigenetic marks, to ensure comparable and harmonized analyses within BovReg.

In terms of data types, our survey allowed us to conclude that most partners are proficient in the analysis of transcriptome data (RNA-Seq), but there is a lack of experience in the analysis of epigenetic data (ChIP-Seq, ATAC-Seq and Hi-C). Therefore, for handling new epigenetic BovReg data sets, specific training sessions should be organized. The results also show that a more in-depth survey regarding pipeline gaps is required. This will be included in the clustering activities related to bioinformatics and pipeline development that are envisaged to be done together with the two other FAANG H2020 projects. As a consequence, the deliverable will be further updated during the course of the project.

2. Introduction

BovReg will provide precise and detailed knowledge of the individual's genetic and epigenetic make-up and its translation into cattle phenotypes. Achieving this will require the integration and analysis of different types of data: phenotypes, genomes/genomic variants, transcriptomes, epigenomes.

In the FAANG community, the need has been discussed to establish genomic annotation pipelines. Their output will be essential to prioritise genetically relevant genomic regions. Furthermore, other types of analyses should be considered, in order to ascertain the biological impact of the identified genomic features at cell biology and phenotype levels. Therefore, it was essential to ascertain the state-of-the-art of BovReg partners' pipelines as well as to identify needs and possible gaps. This information will be used to propose solutions for the issues identified and should constitute a first step towards the final goal of developing reference annotation pipelines, i.e. a set of curated pipelines. They will guarantee the reproducibility of BovReg results as well as their reuse by the community to obtain new annotations when further new data become available.

3. Core report

With the aim of identifying bioinformatics tools and pipelines in use, as well as needs for novel tools, a survey was designed by experts working in WP3. The specific goal of this survey was to identify the tools, which are in use by the different partners, thus allowing the detection of gaps and weaknesses. The draft survey was developed by FMV as the task leader with the input of further experts of WP3. This group included the following persons, Andreia Amaral (FMV-ULisboa), Jose Espinosa-Carrasco and Cedric Notredame (CRG) and Daniel Fischer (LUKE).

3.1 Methods

The survey was designed including 11 sections:

- Section 1- Email address and BovReg Work package (WP).
- Section 2- Workflow information. In this section, the survey aimed to enquire regarding the operationalization of the analysis workflows.
- Section 3- Evaluation of NGS data quality and pre-processing.
- Section 4- Enquire regarding the usage of read mappers while analysing the different types of NGS data.
- Section 5- In this section information is enquired regarding the type of NGS data the participant aims to focus on. After selecting a type of NGS data, the participant is guided to specific survey questions. After answering the specific questions, the participant may choose to return to section 5 again or to submit the survey.
- Section 6- In this section the survey enquires participants regarding the usage of ChIP-Seq quality metrics.
- Section 7- In this section questions regarding Peak calling for ChIP-Seq, ATAC-Seq and Hi-C data were formulated.
- Section 8- This section focused on tools available for post-processing of peaks identified while analysing ChIP-Seq, ATAC-Seq and Hi-C data.
- Section 9- This section was dedicated specifically for Hi-C data.
- Section 10- This section comprises questions regarding different analysis steps of RNA-seq data.
- Section 11- As BovReg aims for the integration of information deriving from genomic features with phenotype and genotype, this last section focused on inquiring regarding the usage of tools to perform genome-wide association testing.

Questions were formulated to obtain multiple-choice responses and all of them also include the following options:

- Section 2 to section 10 : a) I do not know; b) I never did this; c) other;
- Section 11: a) other; b) I am not going to make this type of analysis.

In total, the survey was composed of 36 questions organized into 11 sections. The Google forms platform was used to implement the survey. A private link was sent to BovReg partners working in WP1 to WP7 corresponding to BovReg's scientific work packages. The survey was open from the 15th of January to the 13th of February 2020.

3.2 Results

The survey was answered by 13 BovReg partners who are involved in data analyses.

In regards to participation in WPs (enquired in **Section 1**), most participants are involved in more than one WP (Figure 1), showing the need to integrate different types of NGS (Next Generation Sequencing) data.

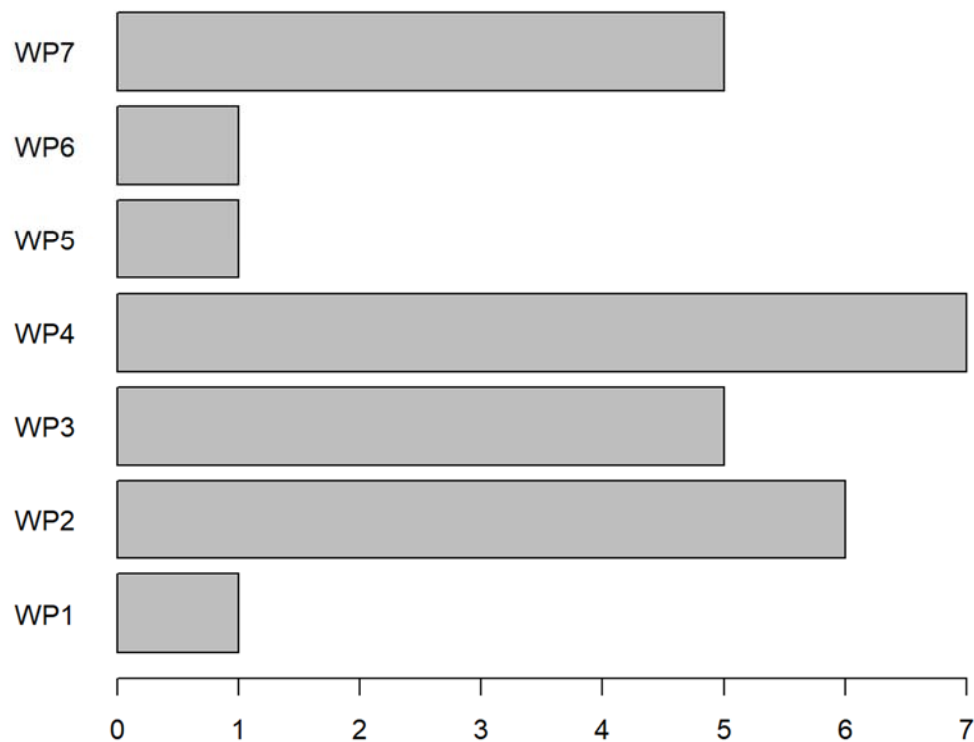


Figure 1: Involvement in BovReg WPs. Bars represent the number of responses of partners who are involved in data analysis in the different scientific WPs.

3.2.1 Section 2

In **Section 2**, the survey aimed to ascertain how users are operating the analysis workflows.

The following questions were formulated:

3.2.1.1 Which of the following workflow management systems do you use?

A total of 4 responses shows that some partners do not use workflow management systems and 3 other affirm that they only use BASH to implement their analyses, while the remaining use either BASH, Snakemake¹ or Nextflow² and one only uses eHive³ (results not shown).

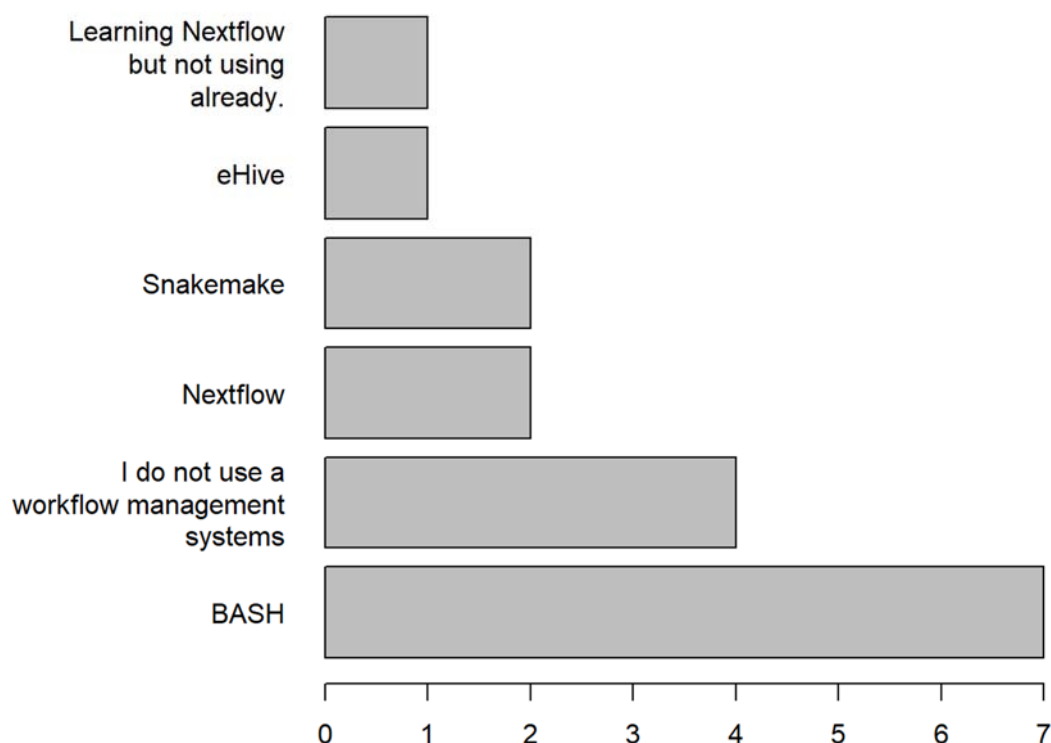


Figure 2: Counts of multiple-choice selection regarding the use of workflow managers. Bars represent the number of times each option was selected by the participants. Options were: a) Snakemake¹; b) Nextflow;² c) BASH; d) I do not use a workflow management systems. e) Other.

3.2.1.2 Which of the following environment management systems do you use?

In Figure 3 it is shown that 6 persons use any of the environment management proposed in the survey, while 7 partners affirm that they do not use environment management systems at all (in particular partners involved in WP2 and WP4, results not shown). From the partners using management systems, half of them prefer containers while half of them use environment modules (3 in each case).

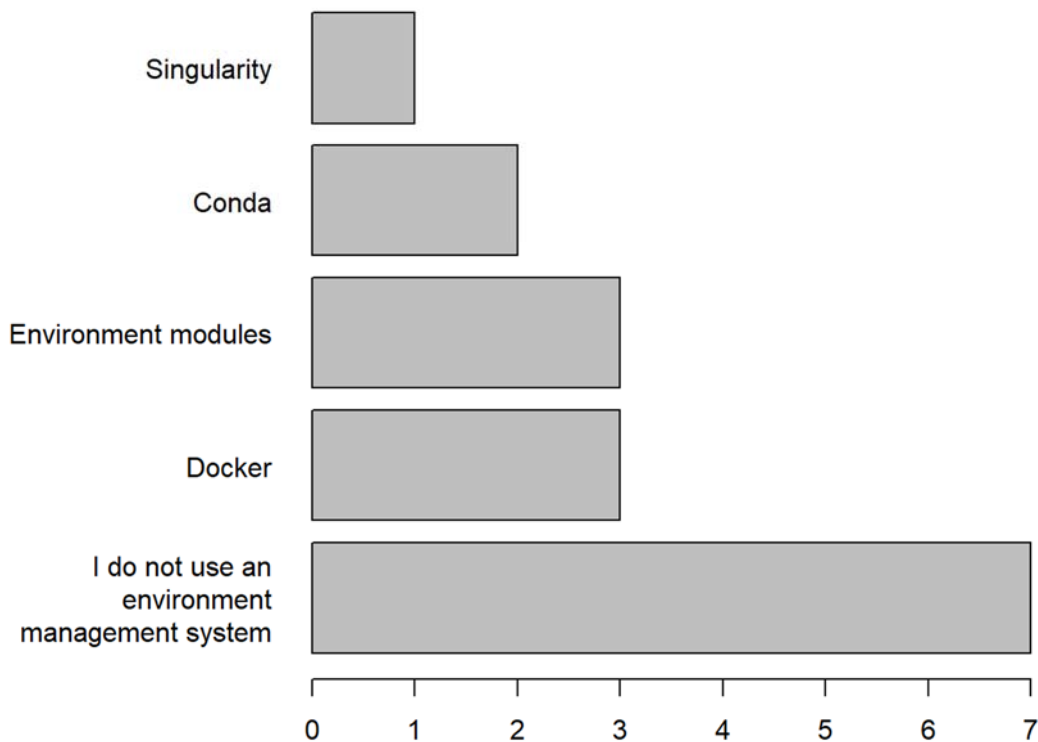


Figure 3 - Counts of multiple-choice selection regarding the usage of environment management systems. Bars represent the number of times each option was selected by the participants. Options were: a) Conda⁴ ; b) Docker⁵; c) Singularity⁶; d) Environment modules; e) I do not use an environment management system; f) Other.

3.2.1.3 Do you use any proprietary software in your pipeline?

BovReg partners declared mostly the use of free license software. .The only proprietary software reported was the Ingenuity Pathway Analysis® software that is used for performing networks and pathway analysis.

3.2.1.4 Where do you usually host your code?

The obtained responses show that most individuals use several different procedures for code storage (Figure 4).

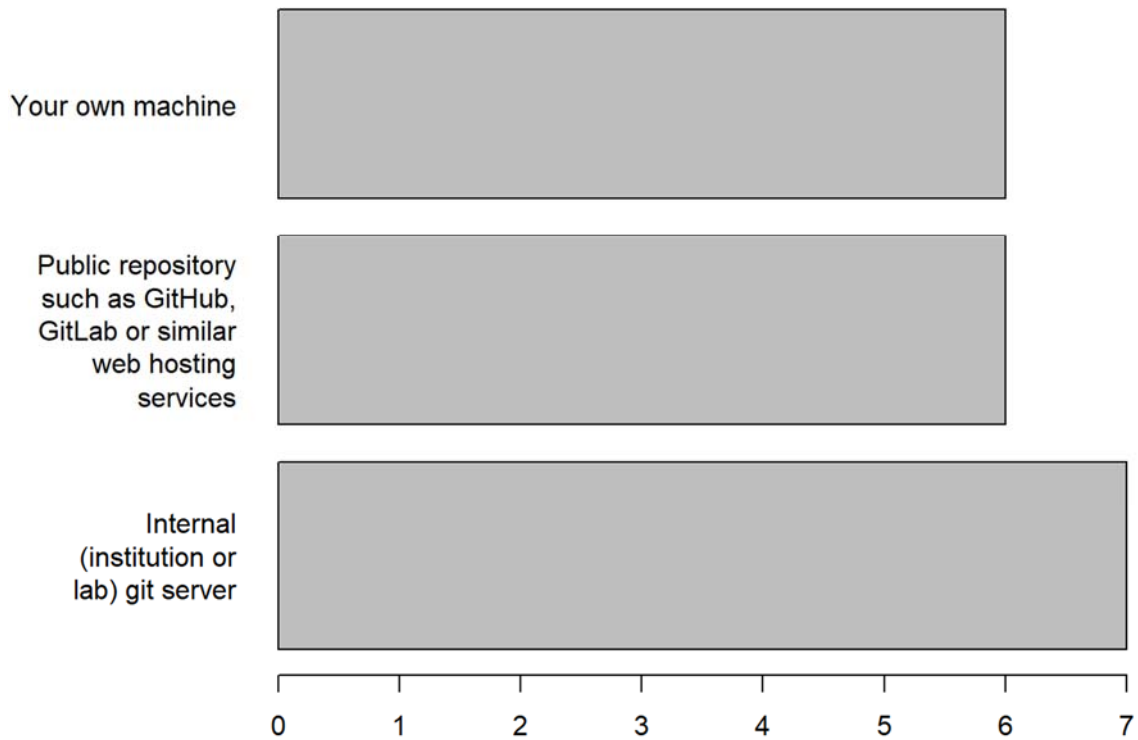


Figure 4: Procedures for code storage. Bars represent the number of times each option was selected by the participants. Options were: a) Public repository such as GitHub⁷, GitLab⁸ or similar web hosting services; b) Internal (institution or lab) git server; c) Your own machine.

3.2.1.5 Where do you usually run your analysis?

From the responses obtained, surveyed participants mostly run analysis processes using high-performance computers (Figure 5).

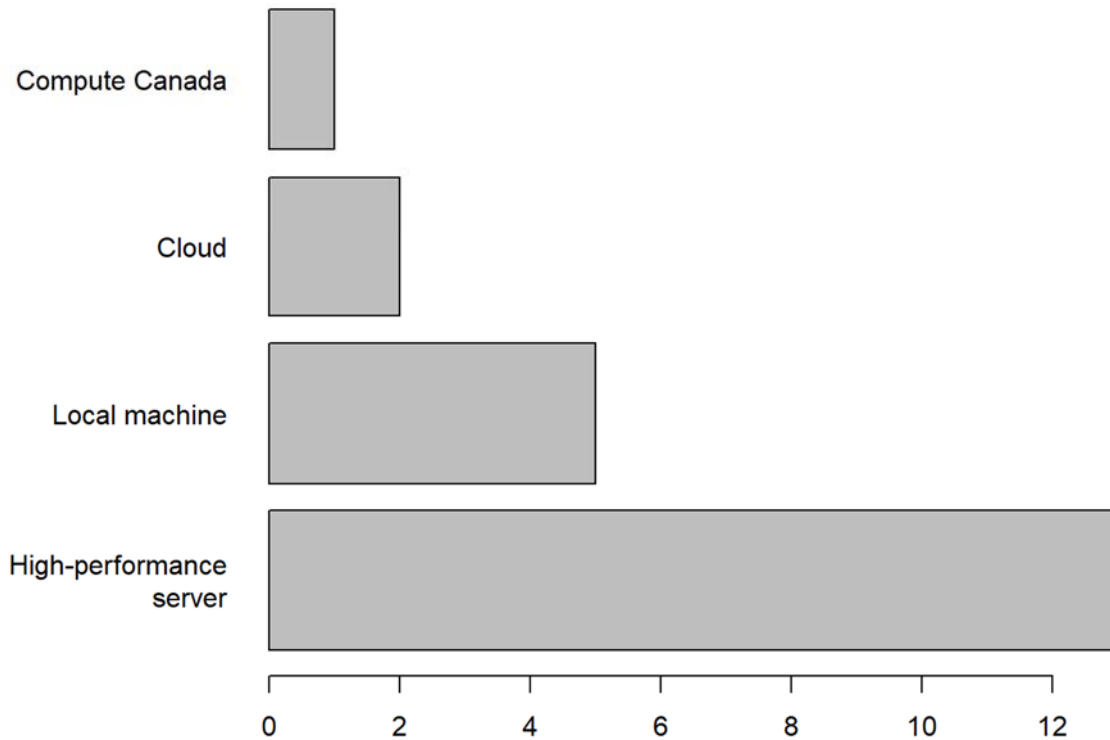


Figure 5 - Informatics facilities for data processing. Bars represent the number of times each option was selected by the participants.

3.2.1.6 On a scale from 1 to 10, how will you rate your pipelines in terms of computational reproducibility (being 10 very reproducible and 1 poorly reproducible?)

About 61% of replies classified the level of reproducibility of their pipelines as very high (8 to 10) while 39% score the reproducibility of their workflows as very low to medium (1 to 6).

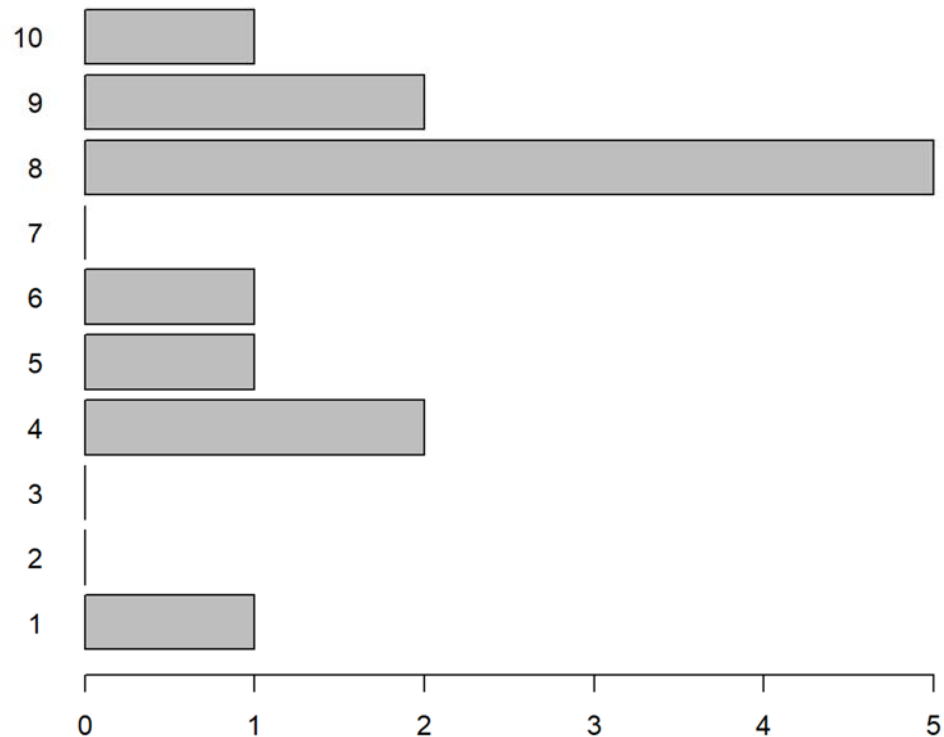


Figure 6: Perception of the level of reproducibility of bioinformatic workflows. Bar charts represent the frequency of selection of the user's perception, ranked 1-10, with 1 for "poorly reproducible" workflows up to 10 "highly reproducible" workflows.

3.2.2 Section 3

A workflow of NGS data processing should start with the implementation of procedures for the evaluation of NGS data quality and it should include pre-processing steps in order to discard low-quality data. Therefore, in **Section 3** of the survey, we have designed questions that aim to evaluate the application of these practices.

3.2.2.1 Which of these tools do you use for evaluation of data quality?

In this question, we have enquired regarding the usage of the following tools for the evaluation of data quality: a) FASTQC⁹; b) Prinseq¹⁰; c) MultiQC¹¹; d) Own scripts; e) Do not know; f) Never did this; e) Other.

Results in Figure 7 show that FASTQC⁹ is the most frequently used tool, but a large proportion of the surveyed partners also uses their own scripts. Most likely due to the fact that FASTQC main quality features have been designed for genomic DNA data, thus suggesting the need to develop specific measures of data quality for ChIP-seq, ATAC-seq, RNA-seq and Hi-C data.

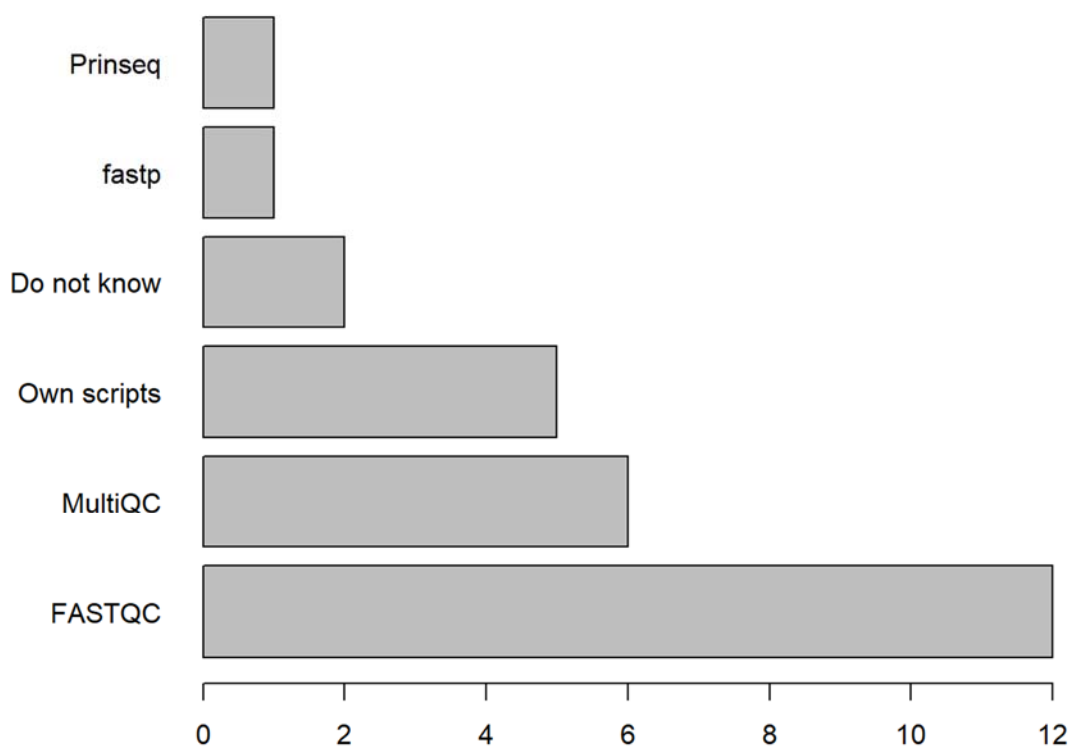


Figure 7-Tools used for evaluation of data quality. Bars represent the number of times each option was selected by each the participants.

3.2.2.2 Which tools do you use for adapter trimming?

Here again, a list of the most commonly used tools was considered for a multiple response answer: a) Cutadapt¹²; b) Flexbar¹³; c) NGmerge¹⁴; d) Trimmomatic¹⁵; d) Own scripts; e) Do not know; f) Never did this; e) Other.

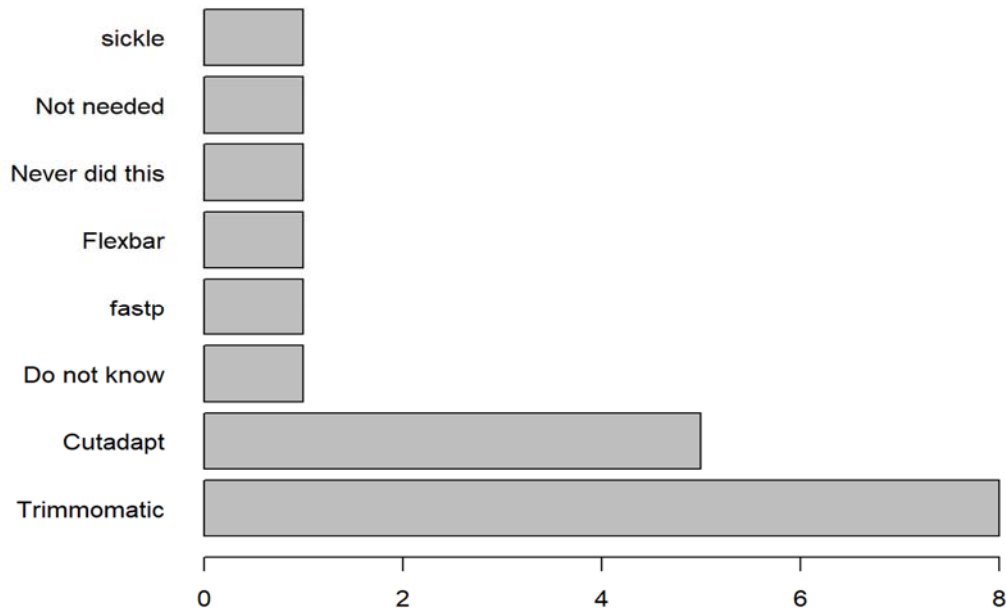


Figure 8- Practical usage of tools for trimming sequencing adapters. Bars represent the number of times each option was selected by the participants.

The obtained results (Figure 8) show that Trimmomatic¹⁵ is the most used tool, followed by Cutadapt¹². Both these tools perform well with datasets for which the sequence of the adapter is known. It is of notice that two of the surveyed respondents answered that they never have removed adapters or that they do not need this step for the analysis.

3.2.2.3 Which tools do you use for quality filtering?

The following options were included according to experts own experience and solutions found in the literature regarding the usage of quality filtering: a) Prinseq¹⁰; b) Flexbar¹³; c) Own scripts; d) Do not know; e) Never did this; f) Other.

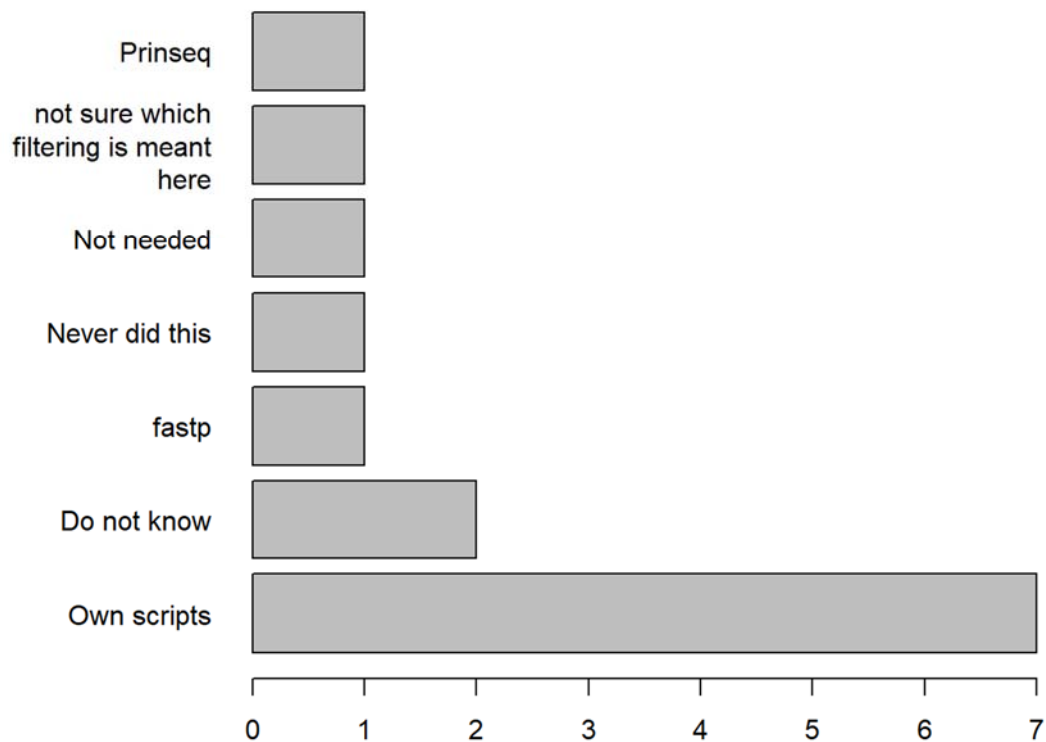


Figure 9 - Usage of quality filtering tools. Bars represent the number of times each option was selected by the participants.

From the obtained responses (Figure 9), it can be observed that for quality filtering most respondents preferably use their own scripts. This suggests that available tools do not allow setting ad-hoc filtering parameters for the type of analyses being performed.

3.2.2.4 Which tools do you use to filter out duplicate sequences?

The possible answers for this question were: a) Picard¹⁶; b) samtools¹⁷; c) Own scripts; d) Do not know; e) Never did this; f) Other.

The obtained results (Figure 10) show that most respondents use Picard.

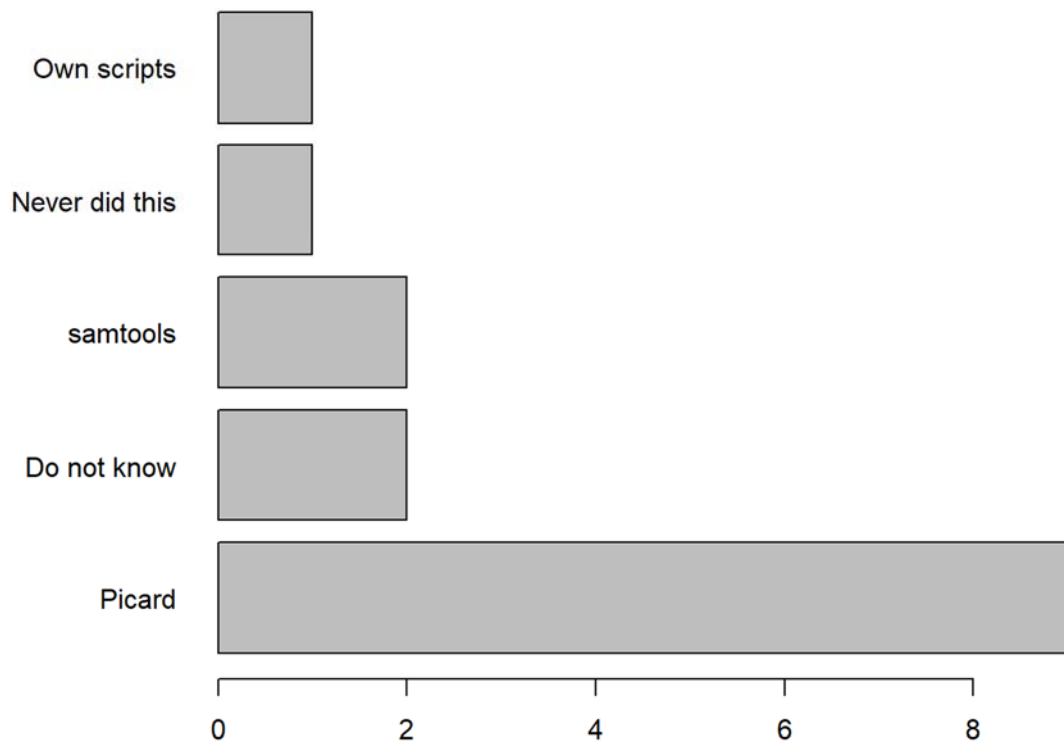


Figure 10: Usage of tools to eliminate duplicate sequences from data. Bars represent the number of times each option was selected by the participants.

3.2.3 Section 4

After this question, the surveyed partners were directed to **Section 4** of the survey in which questions relate to the usage of read mappers for analysing different types of NGS data.

3.2.3.1 Please select the read mappers that you use for each type of data.

To answer this question the participant was required to select at least the analysis of one type of data (RNA-seq, ChIP-Seq, ATAC-seq or Hi-C). These options correspond to the types of NGS data that BovReg aims to generate. Then, for each of these types of data, the following options for read mappers were made available for multiple selection: a) BWA¹⁸; b) Bowtie¹⁹; c) Bowtie2²⁰; d) GSNAP²¹; e) HISAT²²; f) HISAT2²³; g) MAQ²⁴; h) MapSplice2²⁵; i) RUM²⁶; j) Soap²⁷; l) STAR²⁸; m) Tophat2²⁹; n) other; o) Don't analyse this data.

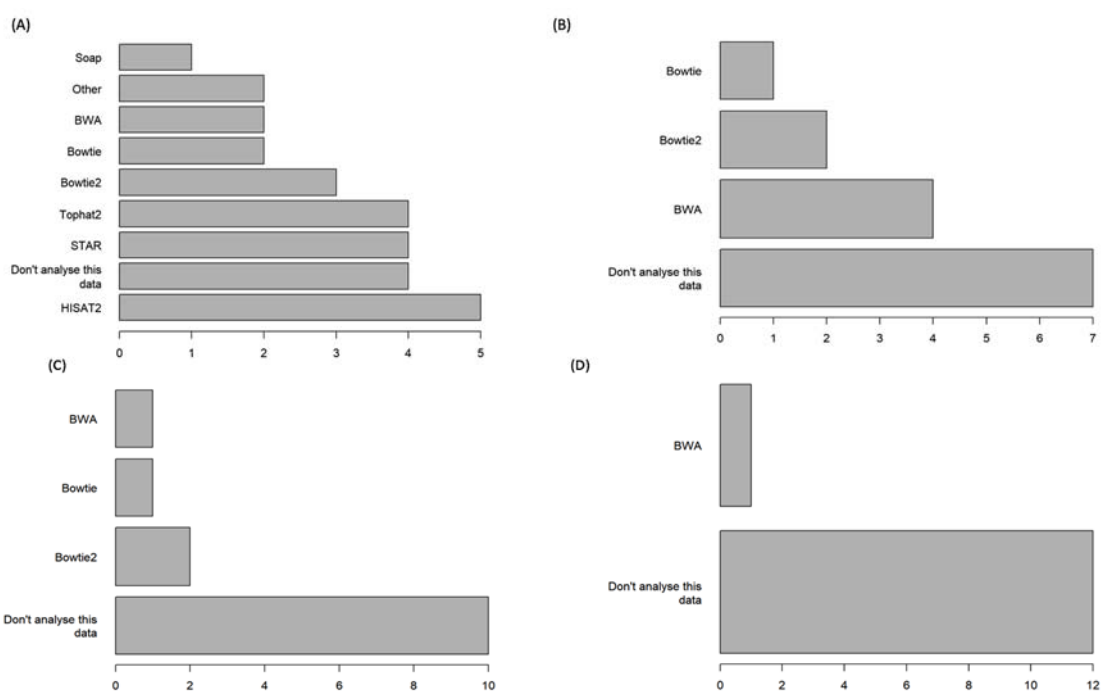


Figure 11: Usage of read mappers. (A) Read mappers used for RNA-seq data; (B) Read mappers used for ChIP-Seq data; (C) Read mappers used for ATAC-Seq data; (D) Read mappers used for Hi-C data. Bars represent the number of times each option was selected by the participants.

3.2.3.2- Which of the tools do you use to check for insert sizes?

As shown in Figure 11, for RNA-seq data, participants have selected multiple read mappers. Some of the selected read mappers do not allow to split reads, and are not adequate for RNA-seq analysis (namely Bowtie and BWA); most likely respondents have selected these while considering miRNA-seq as RNA-seq. Then regarding ChIP-Seq, ATAC-seq and Hi-C data, the most frequent response was that participants have never analysed this type of data.

Respondents that selected option (n) “other” were asked, which tool they use. Two answers were obtained this way and respondents reported that they use Kallisto and Salmon (data not shown).

Checking insert sizes is important as a quality assessment of the data obtained and to infer if the data are mapped to a reference genome in accordance with the experimental design. Although this procedure is very frequent while analysing RNA-seq data, currently there are not many open-source tools for performing this analysis. Therefore the respondents could only select among the following choices: a) BWA¹⁸; b) Picard¹⁶; c) Never did this; d) Don't know.

As shown in Figure 12, BWA and Picard are the most used tools. Four participants have never performed this analysis and two of them answered “do not know”.

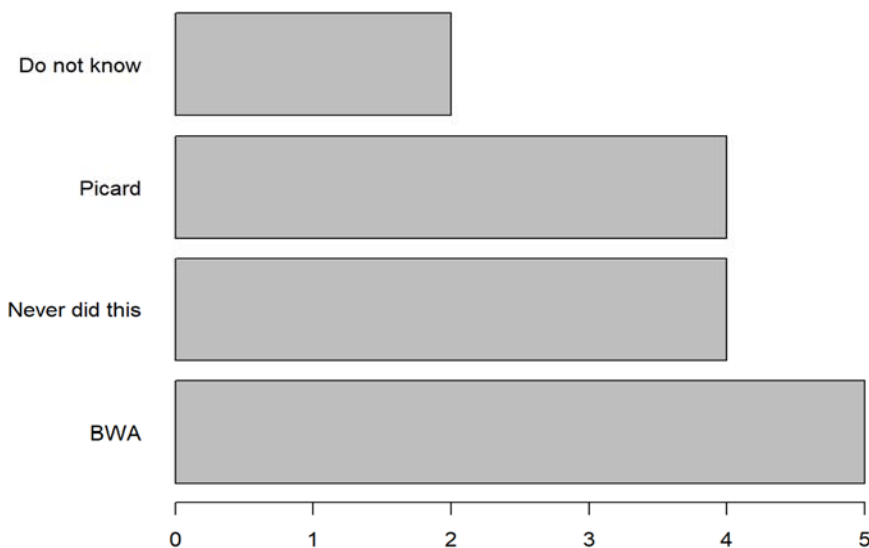


Figure 12 – Enquiring regarding the usage of tools to estimate the size of insert sizes. Bars represent the number of times each option was selected by participants.

The survey form leads participants to **Section 5** after filling in the general questions, which apply to all partners involved. In section 5, they should be directed to survey sections containing questions specific to each type of data after selecting one of the following options. Provided options were: a) ChIP-Seq, ATAC-Seq and Hi-C data; b) RNA-seq; c) GWAS; d) I would like to submit the survey. After selecting one the options, at the end of the respective sections, participants were again directed to section 5, where they could select another option or they could select to submit the survey.

A total of four participants selected option “a” and were directed to **Section 6** that contained questions regarding the assessment of quality metrics of ChIP-Seq data.

3.2.4 Section 6

3.2.4.1- To estimate the fraction of mapped reads that fall into peak regions which tools do you use?

The multiple-choice answers were: a) CHiLin³⁰; b) Own scripts; c) Do not know; d) Never did this; e) Other.

As shown in Figure 13, among the four obtained responses, two use their own scripts, and three use other tools, namely DiffBind³¹, deepTools³², ChIPSeeker³³ or their own scripts. While DiffBind and deepTools in fact provide an estimator for the number of mapped reads falling into a peak called FRiP (which stands for Fraction of Reads in Peaks), ChIPSeeker by itself does not allow estimating this parameter. It allows estimating the coverage of peak regions over chromosomes that afterwards may be used as input to estimate the fraction of mapped reads that fall into peak regions.

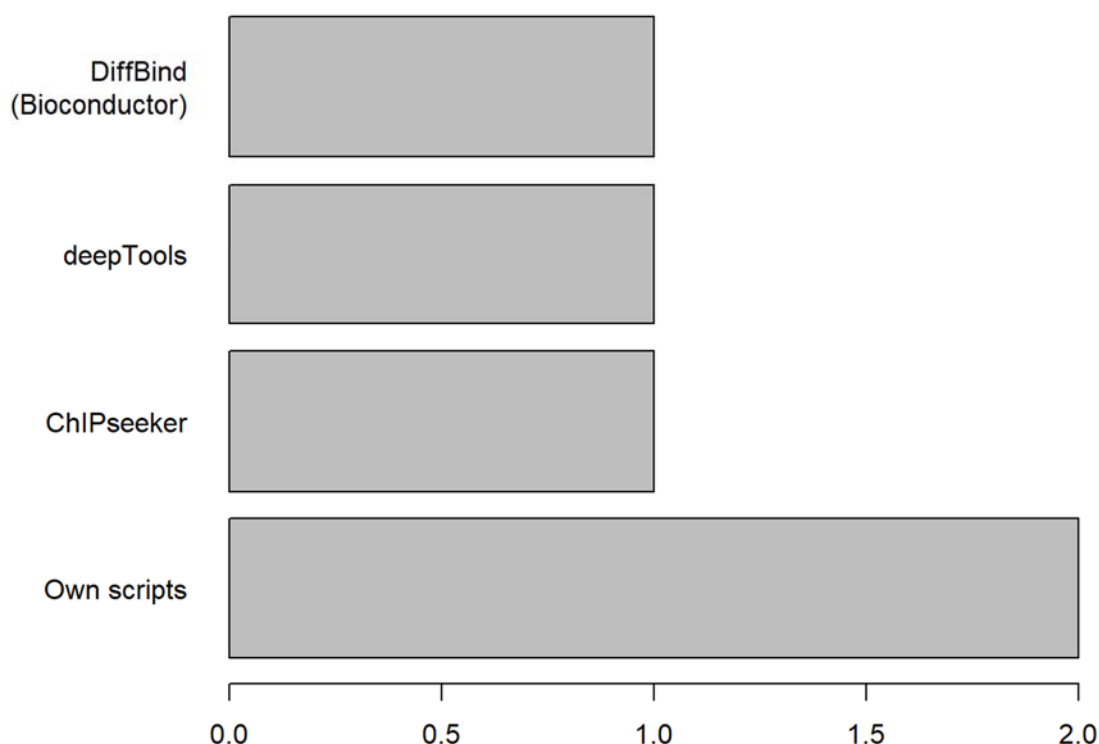


Figure 13- Usage of tools to estimate the fraction of ChIP-seq mapped reads that fall into peak regions. Bars represent the number of times each option was selected by participants.

3.2.4.2- To perform cross-correlation analysis (NSC and RSC), which tools do you use?

The next question dealt with the use of other quality parameters, namely performing cross-correlation analysis. The multiple-choice options were: a) CHiLin³⁰; b) CHANCE³⁴; c) PhantomPeakQualTools³⁵; d) Own scripts; e) Do not know; f) Never did this; g) Other.

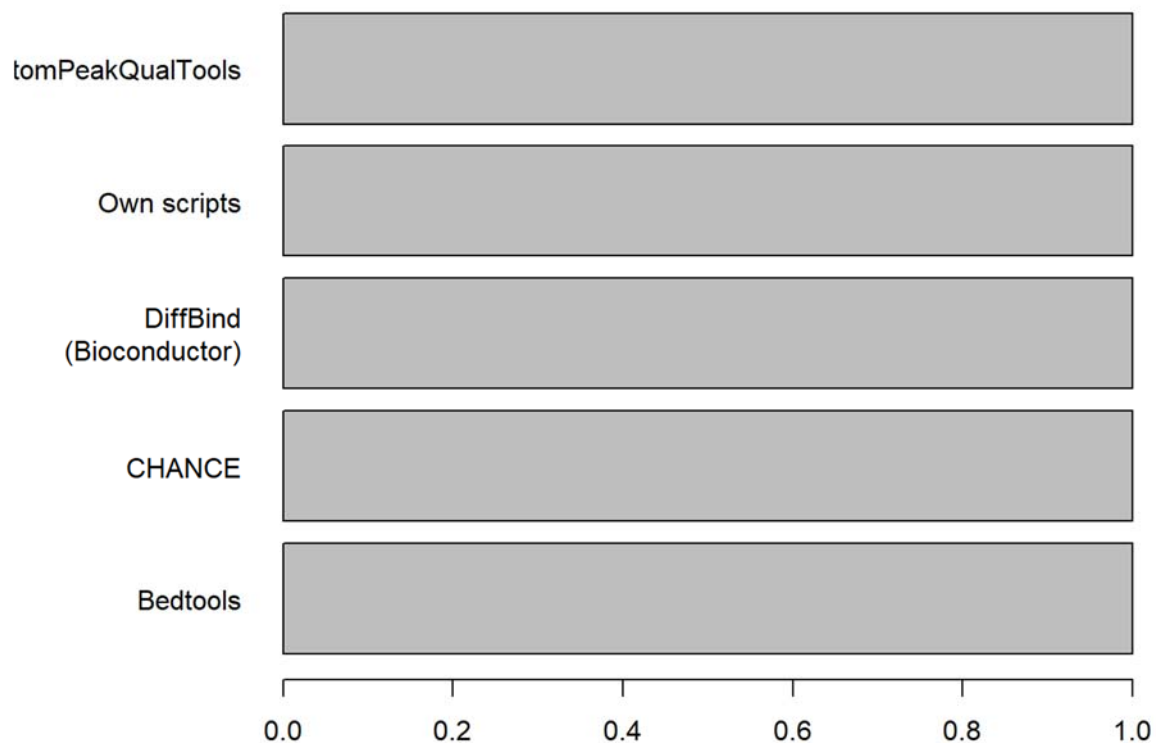


Figure 14- Usage of tools to perform cross-correlation analysis. Bars represent the number of times each option was selected by participants.

Two of the partners use two of the enquired tools, namely CHANCE³⁴ and PhantomPeakQualTools³⁵. Own scripts are also used as well as other software such as Bedtools³⁶ and DiffBind³¹. As BedTools³⁶ does not allow to perform cross-correlation analysis directly and hence, this tool was not considered as an option. Nevertheless, BedTools' users may estimate the variables required to perform cross-correlation analysis. Regarding DiffBind³¹, this tool neither allows to perform this analysis nor to estimate the required variables.

This question finalizes Section 6. Next, the participants in this survey who selected option "a" were directed to **Section 7**. In this section, they were asked in regard to the tools used for performing Peak Calling while analysing ChIP-Seq data (considering different types of ChIP-Seq assays), ATAC-Seq- and Hi-C data.

3.2.5 Section 7

3.2.5.1 - For ATAC-seq, data which peak caller do you use?

Possible multiple-choice options were: a) Genrich³⁷; b) MACS2³⁸; c) BAMPE; d) Do not know; e) Never did this; f) Other.

Obtained responses (Figure 15), show that for the four partners who replied to the question related to this data category, two have never performed analyses of ATAC-Seq data. For the remaining partners, Genrich and MACS2 were the selected options.

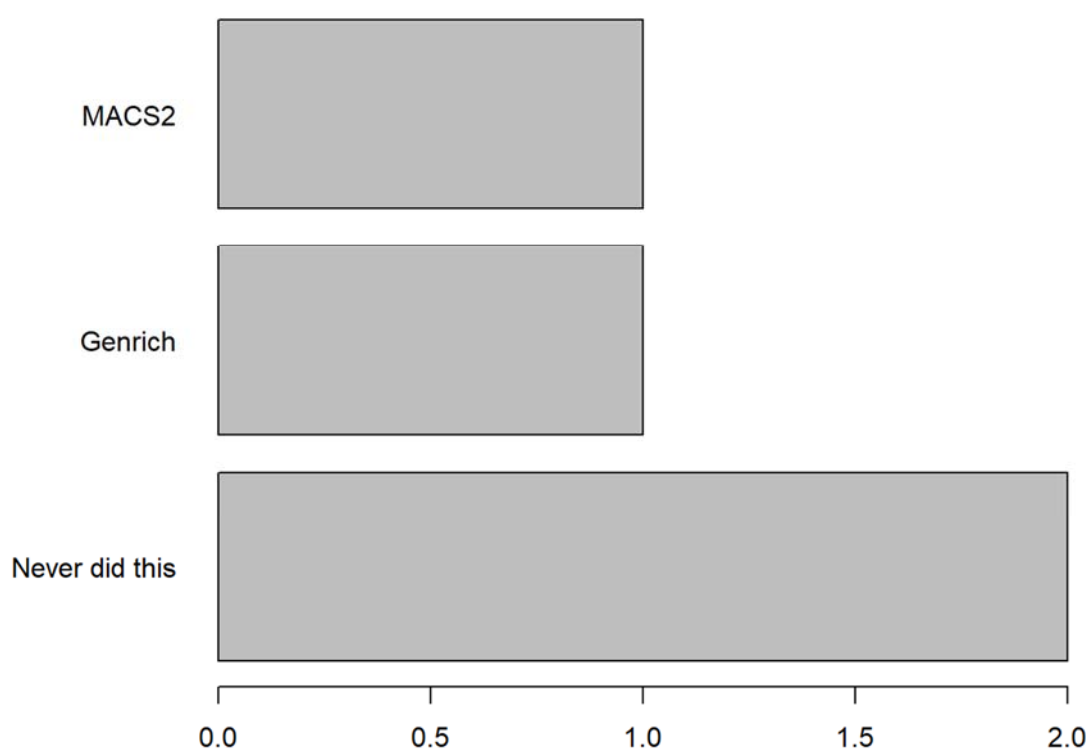


Figure 15 - Usage of tools to perform Peak calling while analysing ATAC-Seq data. Bars represent the number of times each option was selected by participants.

3.2.5.2- For ChIP-Seq (Transcription Factors) data, which peak caller do you use?

Possible multiple-choice options were: a) MACS³⁹; b) MACS2³⁸; c) SIPEs⁴⁰; d) SPP⁴¹; e) CSAR⁴²; f) GPS; g) polyAPeak; h) Do not know; i) Never did this; j) Other.

From the obtained responses (Figure 16), MACS and MACS2 are being used as well as Genrich and SWEMBL. These last two software tools were referred to in the category of “other”.

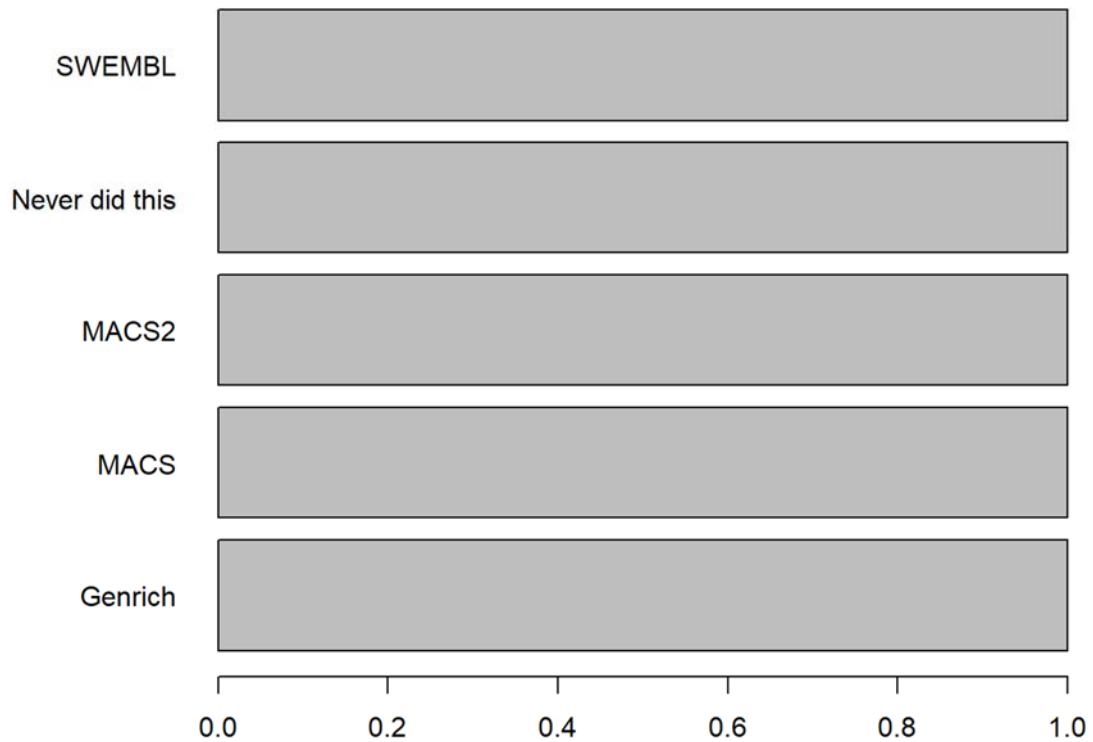


Figure 16- Usage of tools to perform Peak calling while analysing ChIP-Seq data generated to investigate binding sites for Transcription Factors. Bars represent the number of times each option was selected by participants.

3.2.5.3- For ChIP-Seq (Histones) data, which peak caller do you use?

Possible multiple-choice options were: a) SPP⁴¹; b) MACS³⁹; c) SICER⁴³; d) CCAT⁴⁴; e) ZINBA⁴⁵; f) RSEG⁴⁶; g) Scripture⁴⁷; i) Do not know; j) Never did this; l) Other.

Here again, participants in the survey use very different approaches to deal with this data type (Figure 17). From the options given in the question CCAT⁴⁴, MACS³⁹ and SICER⁴³ were selected once as the preferred tool, while other tools considered by participants choosing “other” were identified as Genrich³⁷ and SWEMBL⁴⁸.

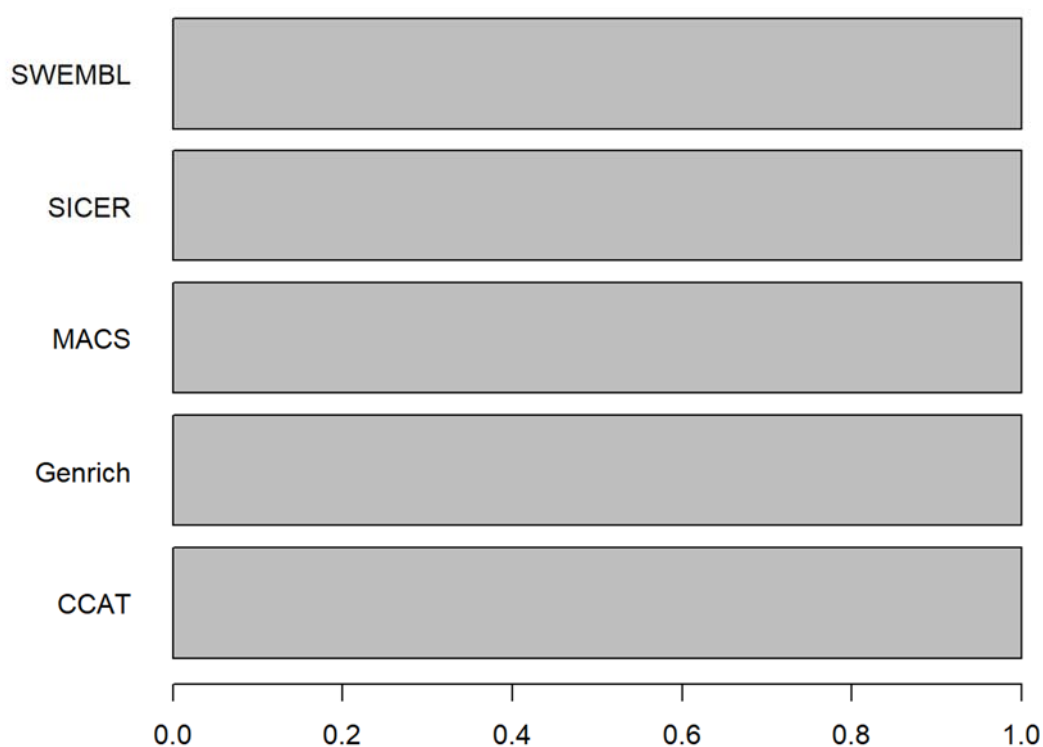


Figure 17 - Usage of tools to perform Peak calling while analysing ChIP-Seq data generated to investigate binding sites for Histones. Bars represent the number of times each option was selected by participants

3.2.5.4- For ChIP-Seq (Pol III) data which peak caller do you use?

Possible multiple-choice options were: a) SPP; b) MACS; c) ZINBA; d) PeakRanger; e) Do not know; f) Never did this; g) Other.

Results (Figure 18) show that only one of the participants had analysed ChIP-Seq data for the detection of Pol III binding sites before. This person reported to use MACS to perform peak calling.

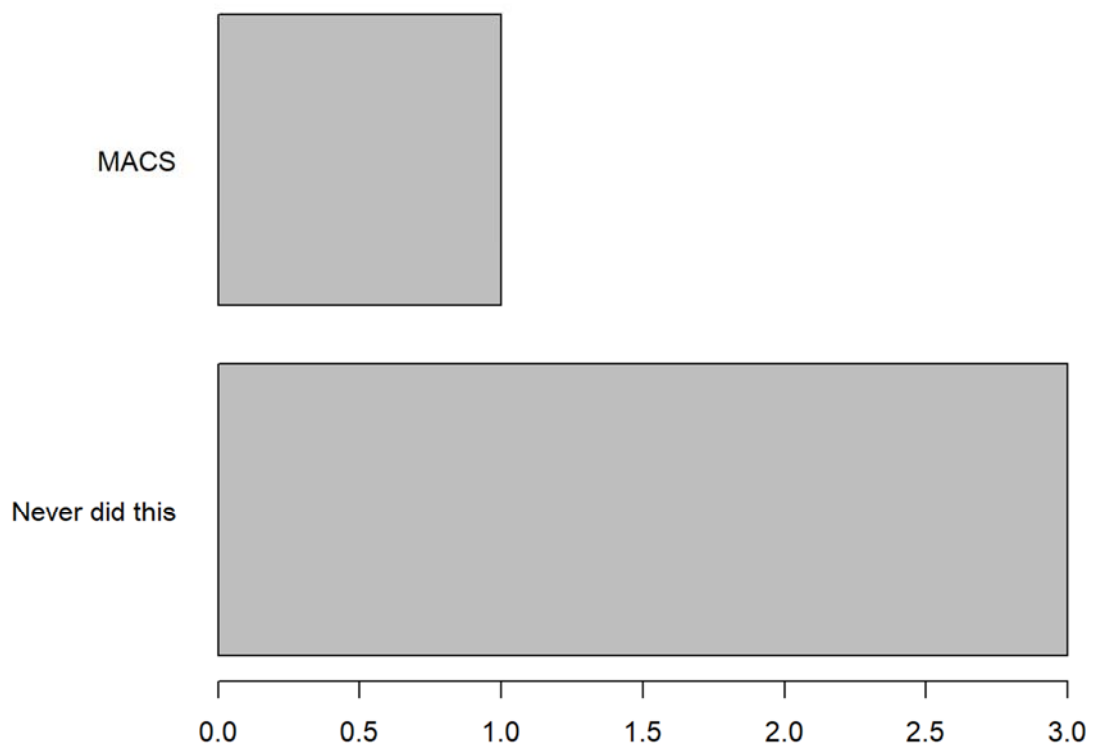


Figure 18 - Usage of tools to perform Peak calling while analysing ChIP-Seq data generated to investigate binding sites for *Pol III*. Bars represent the number of times each option was selected by participants.

After peak calling for ChIP-Seq, ATAC-Seq and Hi-C analyses post-processing of the obtained peaks is needed. Thus, participants were directed to **Section 8** for further questions regarding this issue.

3.2.6 Section 8

3.2.6.1 Which tool do you use for normalisation?

Normalization aims to harmonize read counts among datasets in order to allow performing comparative testing analysis. The included multiple-choice options to perform this step were: a) PeakSeq⁴⁹; b) CisGenome⁵⁰; c) MACS³⁹; d) USeq⁵¹; e) RPKM⁵²; f) POLYPHEMUS⁵³; g) DESeq⁵⁴; h) Do not know; i) Never did this; l) Other.

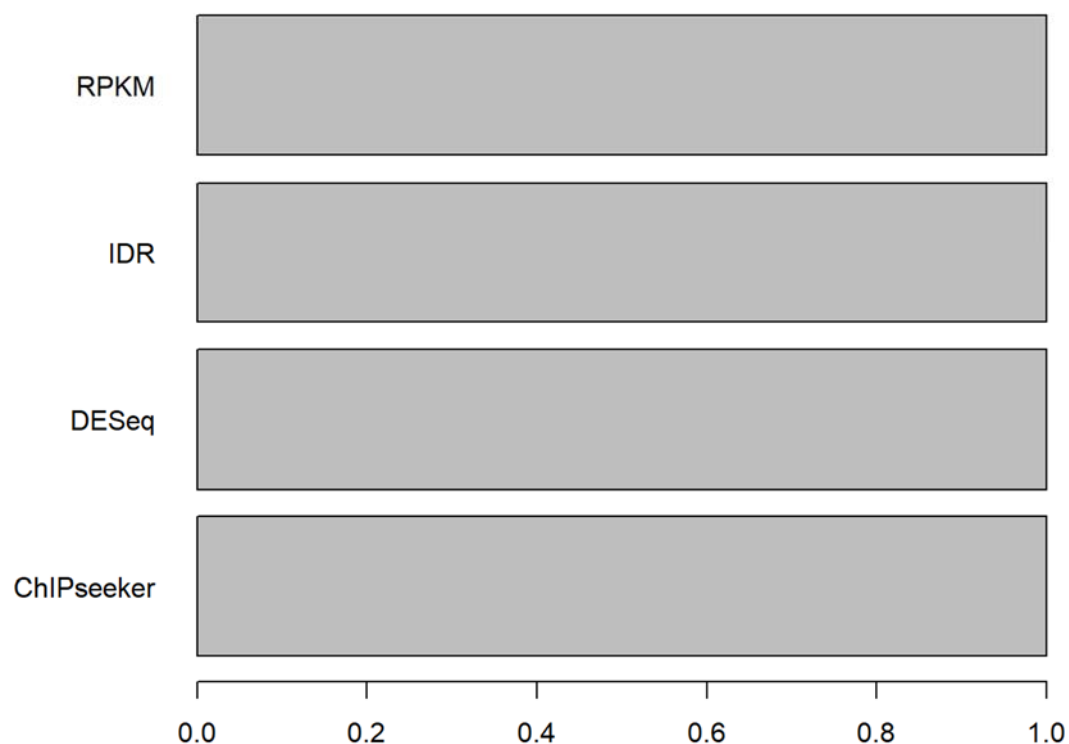


Figure 19 - Usage of tools to perform Read count normalisation while analysing ChIP-Seq, ATAC-Seq and Hi-C data. Bars represent the number of times each option was selected by participants.

Results (Figure 19) show that DESeq⁵⁴ and RPKM methods are being used, however respondents also identified other software namely, ChIPseeker³³ and IDR⁵⁵.

3.2.6.2 Which of these tools do you use for differential binding analysis?

Possible multiple-choice options were: a) BEDTools³⁶; b) DBChIP;⁵⁶ c) MAnorm⁵⁷; d) DIME⁵⁸; e) ChIPDIFF⁵⁹ (histones); f) POLYPHEMUS⁵³ (RNA Pol III); g) Do not know; h) Never did this; i) Other.

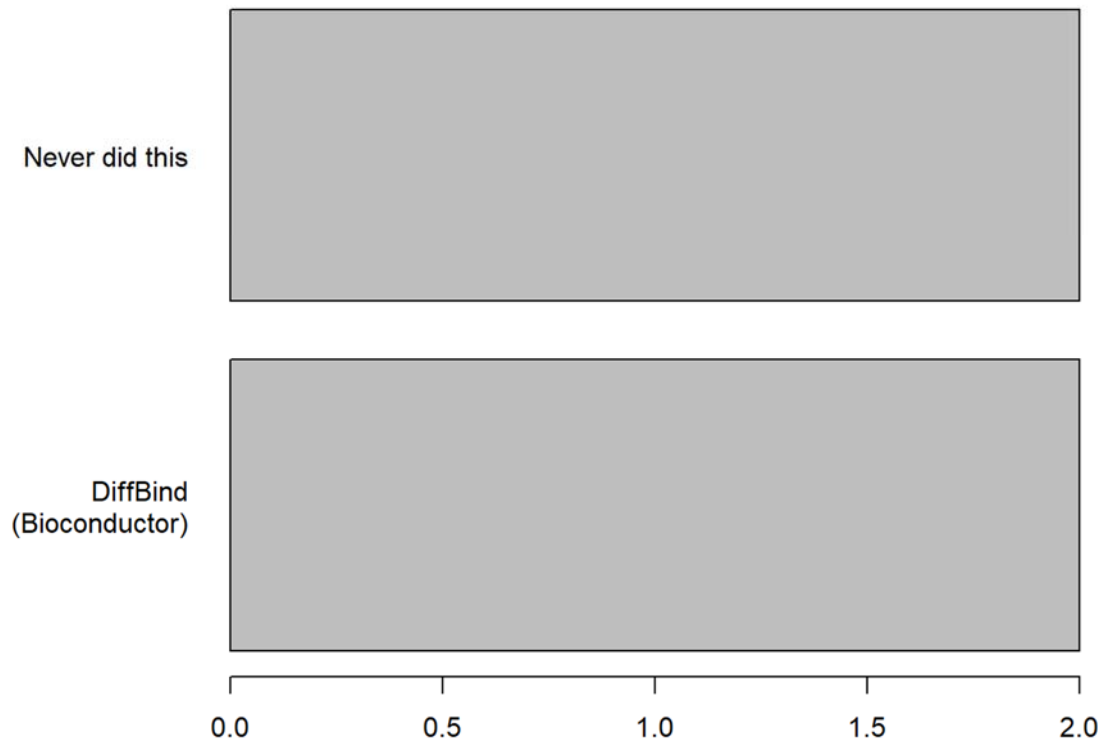


Figure 20 - Usage of tools to perform testing for differences in binding while analysing ChIP-Seq, ATAC-Seq and Hi-C data. Bars represent the number of times each option was selected by participants.

Results (Figure 20) show that half of the interviewed partners never did this analysis. When performing, DiffBind was the only identified tool.

3.2.6.3 Which of these tools do you use for Peak annotation?

Possible multiple-choice options were: a) CEAS⁶⁰; b) ChIPpeakAnno⁶¹; c) Do not know; d) Never did this; e) Other.

Figure 21 shows that respondents do not use any of the tools included in the multiple options. The tools reported under the option “other” were Homer⁶² and ChIPseeker³³.

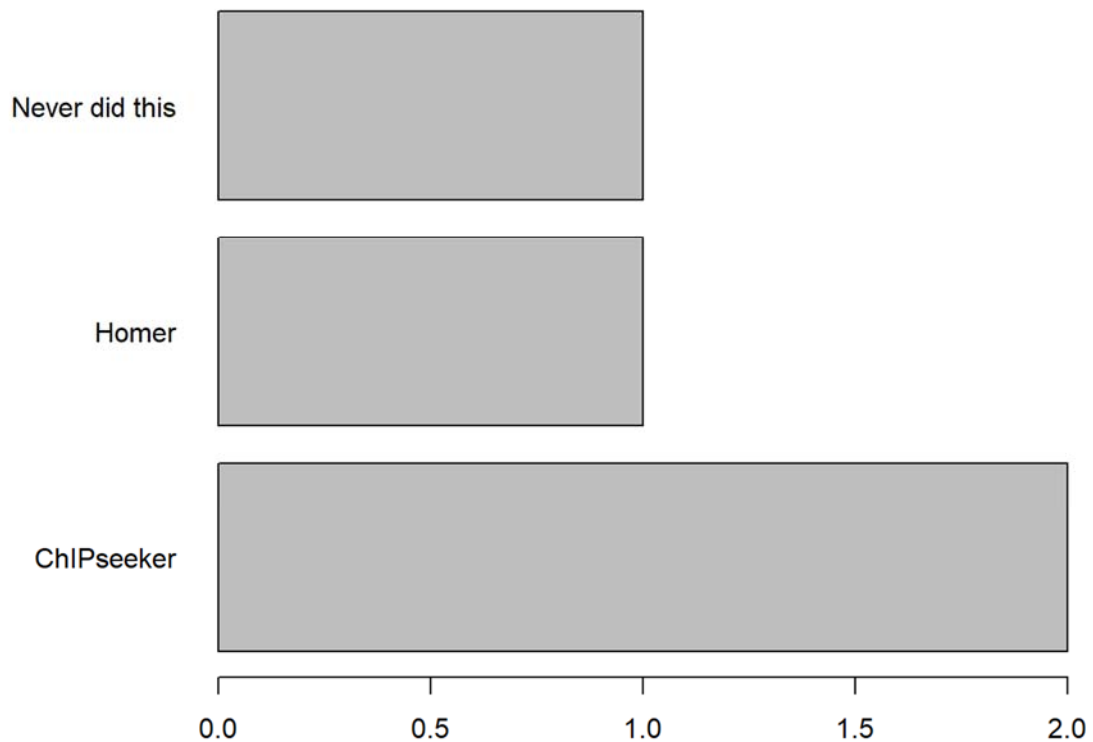


Figure 21 -Usage of tools to perform Peak annotation while analysing ChIP-Seq, ATAC-Seq and Hi-C data. Bars represent the number of times each option was selected by participants.

3.2.6.4 Which of these tools do you use for motif analysis?

Possible multiple-choice options were: a) MEMEChIP⁶³; b) HOMER⁶²; c) Do not know; d) Never did this; e) Other.

MEMEChIP⁶³ and HOMER⁶² were reported as the used options for motif analysis, with the former being preferred over the latter as depicted in Figure 22.

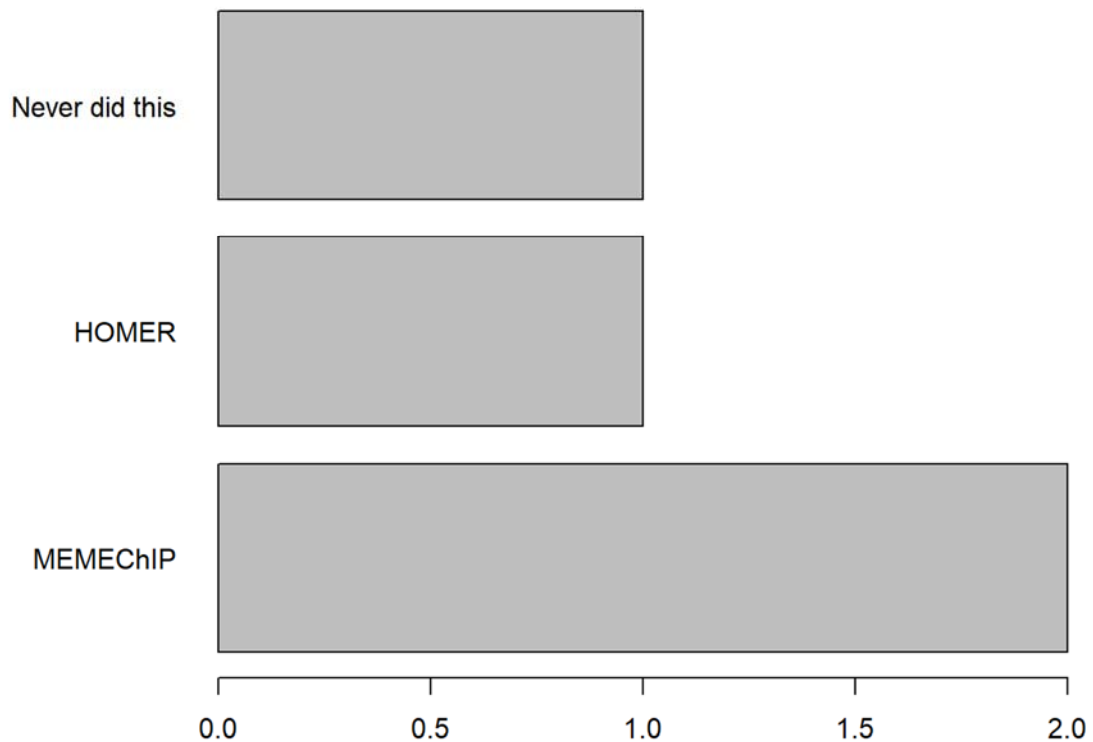


Figure 22 - Usage of tools to perform search for Binding Motifs while analysing ChIP-Seq, ATAC-Seq and Hi-C data. Bars represent the number of times each option was selected by participants.

After this question, respondents were directed to **Section 9** that only included questions regarding the use of pipelines that allow integrative processing of Hi-C data.

3.2.7 Section 9

3.2.7.1 Which tools do you use for integrative processing of Hi-C data?

All respondents answered that they have never performed this analysis. It would be important to clarify whether BovReg partners actually lack this kind of expertise or whether trained people on this analysis type just did not answer the survey.

3.2.8 Section 10

Section 10 of the survey was designed to include questions regarding specific steps of processing and analysis of RNA-seq data. This section included a total of five questions and was filled by seven participants.

3.2.8.1 Which tools do you use to perform gene quantification?

Possible multiple-choice options were: a) HTSeq⁶⁴; b) IRAP⁶⁵; c) RSEM⁶⁶; d) Own scripts; e) Do not know; f) Never did this; g) Other.

Responses obtained (Figure 23) show that half of the respondents choose HTSeq and half select RSEM to perform gene quantification.

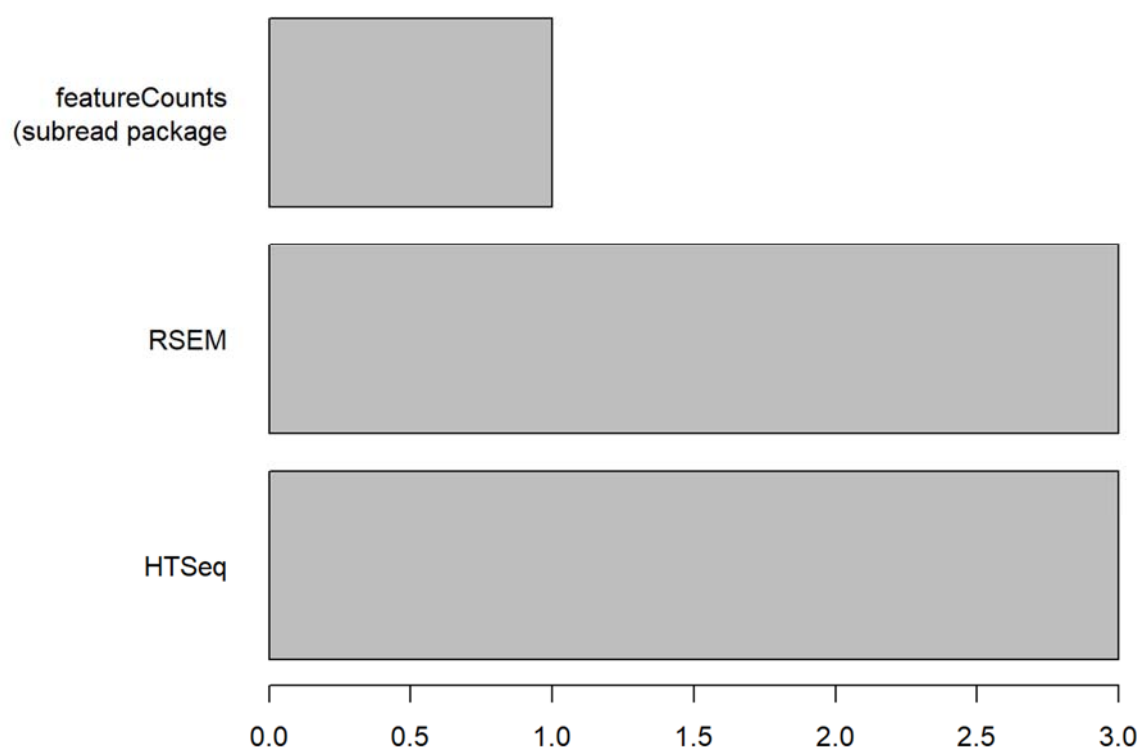


Figure 23 - Usage of tools to perform gene quantification while analysing RNA-Seq data. Bars represent the number of times each option was selected by participants.

3.2.8.2 Which tools do you use for isoform detection and quantification?

Possible multiple-choice options were: a) RSEM⁶⁶; b) eXpress⁶⁷; c) Tigar2⁶⁸; d) BitSeq⁶⁹; e) Kallisto⁷⁰; f) RapMap⁷¹; g) Salmon⁷²; e) Cufflinks⁷³; f) Sailfish⁷⁴; g) Do not know; h) Never did this; i) Other.

As shown in Figure 24, Cufflinks²⁹ and RSEM⁶⁶ were among the most used tools to perform this analysis while Kallisto⁷⁰, Salmon⁷² and Stringtie⁷⁵ utilization was reported just one time each.

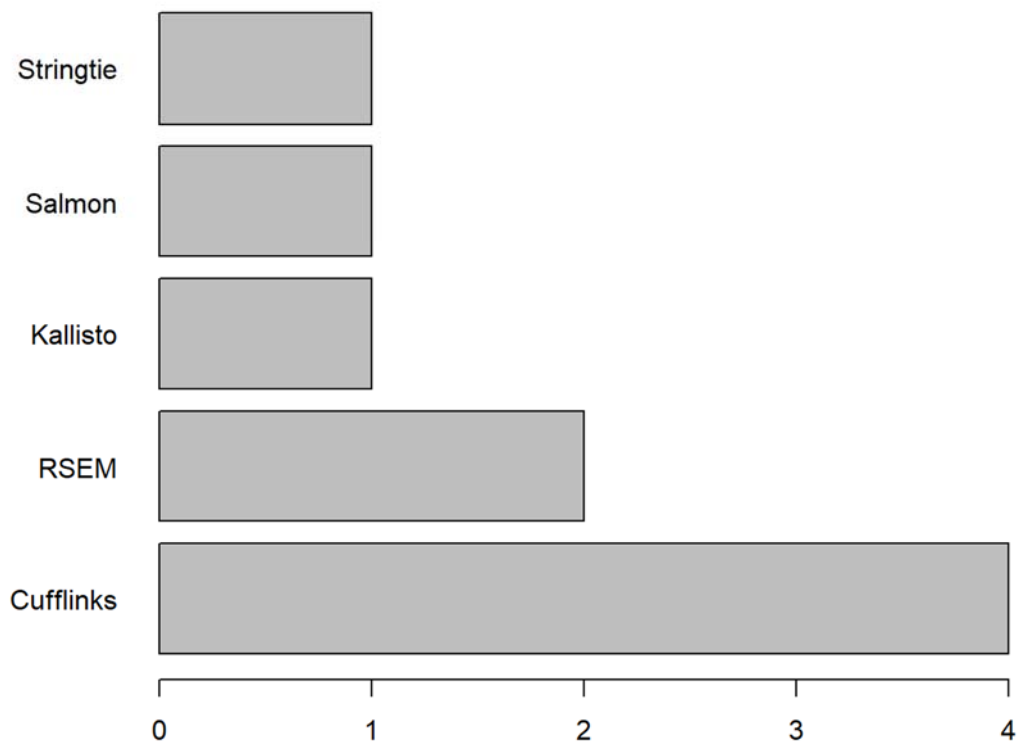


Figure 24 - Usage of tools to perform isoform quantification while analysing RNA-Seq data. Bars represent the number of times each option was selected by participants.

3.2.8.3 Which tool/method do you use for read normalisation?

Possible multiple-choice options were: a) RPKM⁵²; b) FPKM⁷⁶; c) RSEM⁶⁶; d) EdgeR⁷⁷; e) DESeq⁵⁴; f) DESeq2⁷⁸; g) Do not know; f) Never did this; h) Other.

Results (Figure 25) show that participants use several of the tools since most of them have selected multiple options. However, participants preferred DESeq2⁷⁸, FPKM⁷⁶ and EdgeR⁷⁷ over the other options and no other tools besides the ones included in the multiple options were provided.

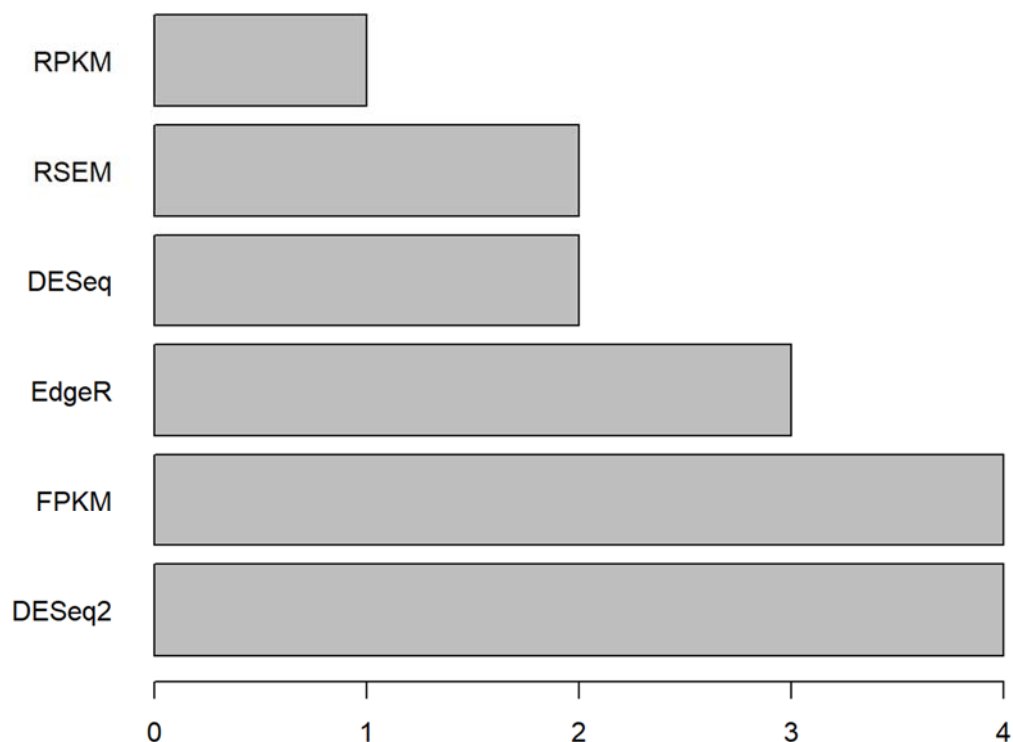


Figure 25 - Usage of tools to perform read count normalisation while analysing RNA-Seq data. Bars represent the number of times each option was selected by participants.

3.2.8.4 - Which tool do you use for differential expression analysis?

Possible multiple-choice options were: a) EdgeR⁷⁷; b) DESeq⁵⁴; c) DESeq2⁷⁸; d) Limma-voom⁷⁹; e) NOISeq⁸⁰; f) Do not know; g) Never did this; h) Other.

DESeq2⁷⁸ was the preferred option to perform differential expression analysis since its use was reported by all the respondents but one. However, EdgeR⁷⁷ and DESeq⁵⁴ seem to be also quite popular among our partners (Figure 26).

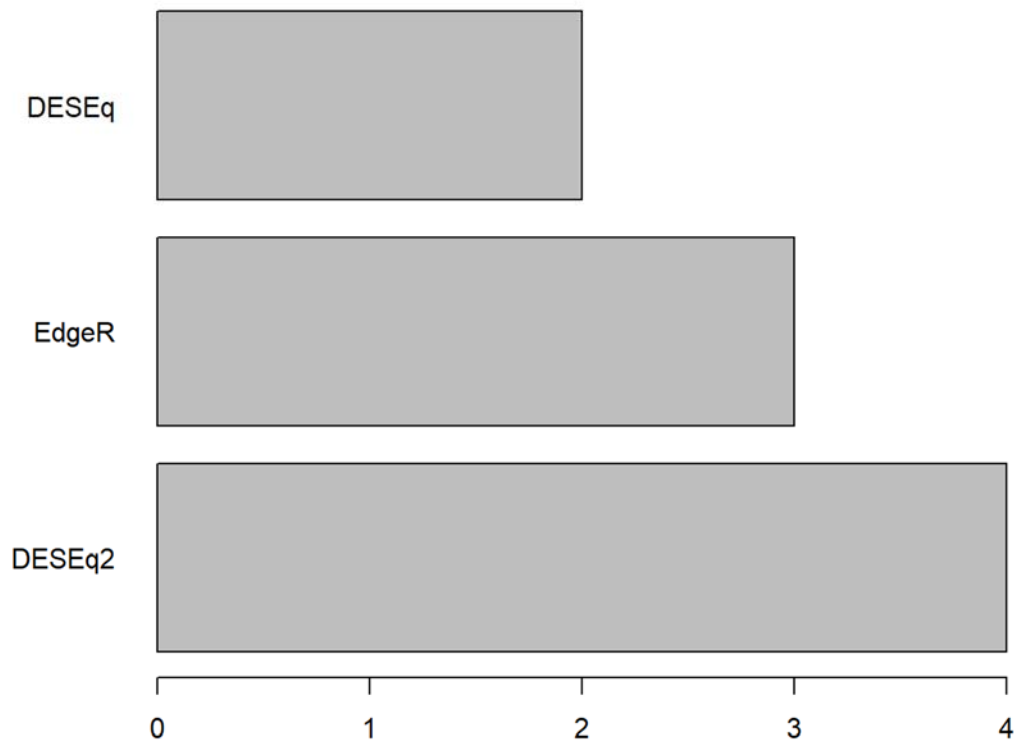


Figure 26 - Usage of tools to perform differential expression testing. Bars represent the number of times each option was selected by participants.

3.2.8.5 Which tool do you use for lncRNA prediction and identification?

Possible multiple-choice options were: a) FEELnc/Flexible⁸¹ extraction of lncRNAs; b) lncRScan-SVM⁸²; c) Do not know; d) Never did this; e) Other.

The obtained results (Figure 27) show that most respondents used FEELnc⁸¹ for the analysis of lncRNA, however other software packages are currently in use by BovReg partners, namely CNCI⁸³, PLAR⁸⁴ and PLEK⁸⁵.

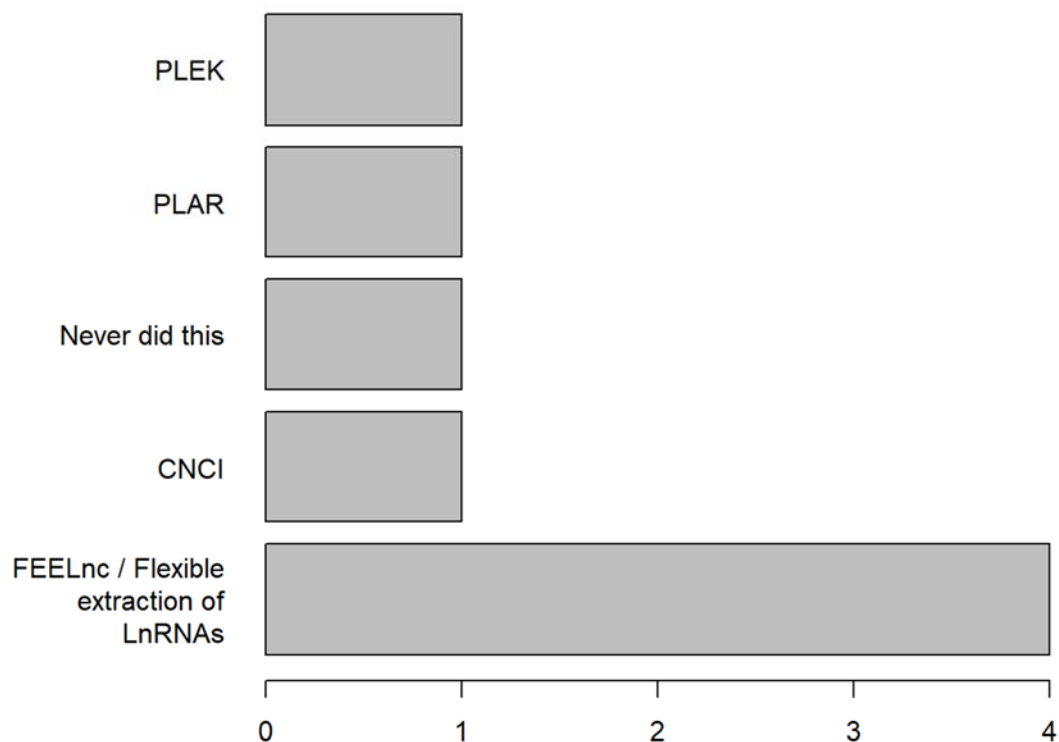


Figure 27 - Usage of tools to perform lncRNA prediction and identification. Bars represent the number of times each option was selected by participants.

Finally, the survey included one section regarding algorithms to perform association testing and quality control of genotyping data. **Section 11** is comprised of four questions. A total of 10 respondents answered this section of the survey.

3.2.9 Section 11

3.2.9.1 For pre-processing of genotype data and QC control, which tools do you use?

Possible multiple-choice options were: a) qctool⁸⁶; b) bcftools⁸⁷; c) vcftools⁸⁸; d) PLINK2⁸⁹; e) GenABEL⁹⁰; f) GWASTools⁹¹; g) Not going to make this type of analyses; h) Other.

Obtained results (Figure 28) show that the preferred option was PLINK2⁸⁹ closely followed by vcftools⁸⁸ and the bcftools⁸⁷ while the use of GenABEL was only reported by one of the partners. Notably, respondents use several of the enquired tools to perform this type of analysis.

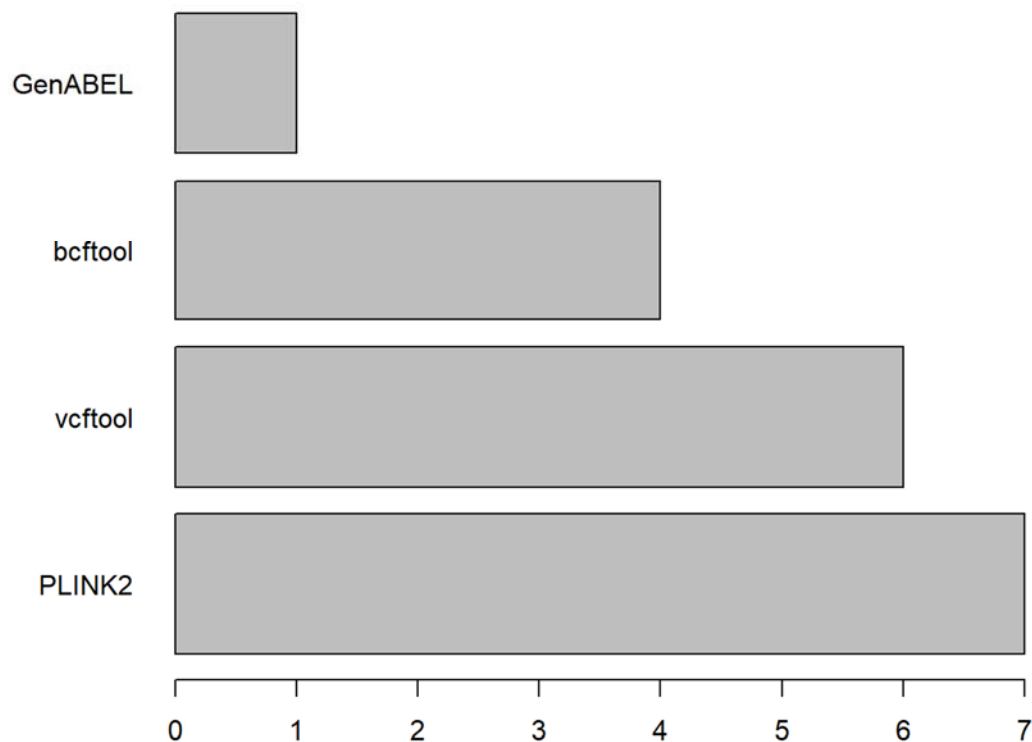


Figure 28 - Usage of tools to perform QC control of genotype data. Bars represent the number of times each option was selected by participants.

3.2.9.2 For phasing, which of these tools do you use?

Possible multiple-choice options were: a) SHAPEIT⁹²; b) Beagle⁹³; c) Unphases⁹⁴; d) Not going to make this type of analysis; e) Other.

Results in Figure 29) show that Beagle is the most used tool, followed by SHAPEIT⁹² and EAGLE⁹⁵.

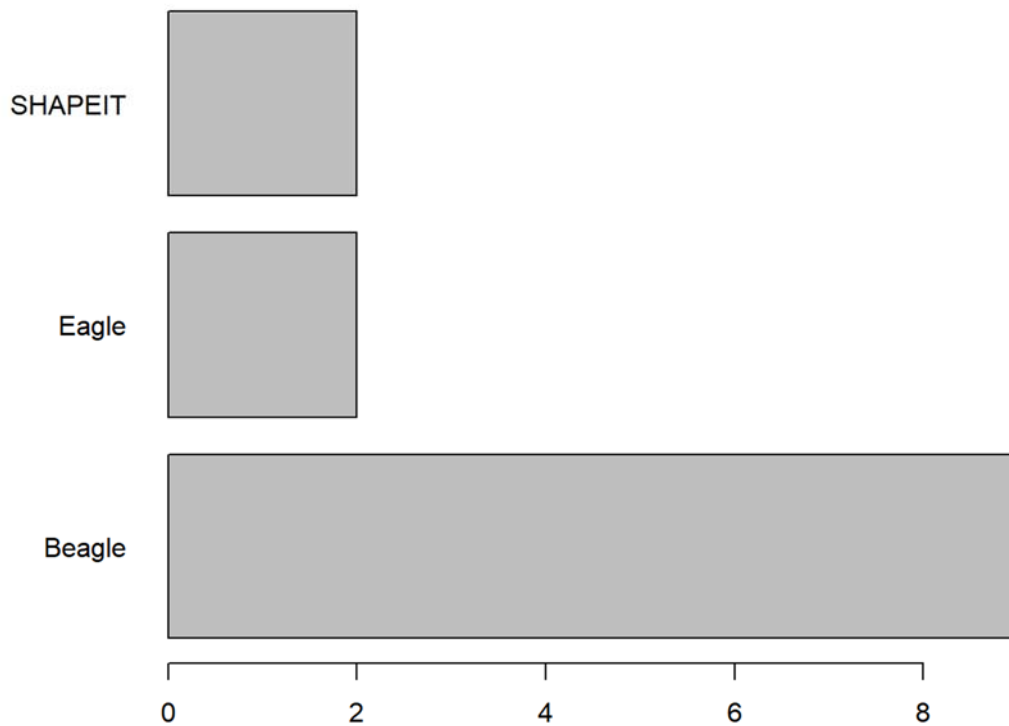


Figure 29 - Usage of tools to perform phasing of quality approved of genotype data. Bars represent the number of times each option was selected by participants.

3.2.9.3 - For imputation which of these tools do you use?

Possible multiple-choice options were: a) IMPUTE2⁹⁶; b) Beagle⁹³; c) MaCH⁹⁷; d) MaCH-Minimac⁹⁷; e) MaCH-Admix⁹⁷; f) PBWT⁹⁸; g) SNPMatrix⁹⁹; h) snpSTATS¹⁰⁰; i) Not going to make this type of analyses; j) Other.

Results show (Figure 30) that although one tool is selected by more respondents, most of the participants selected multiple tools, showing that they are not restricted to the use of a single tool to perform this analysis. Again, the most popular tool is Beagle closely followed by IMPUTE2⁹⁶.

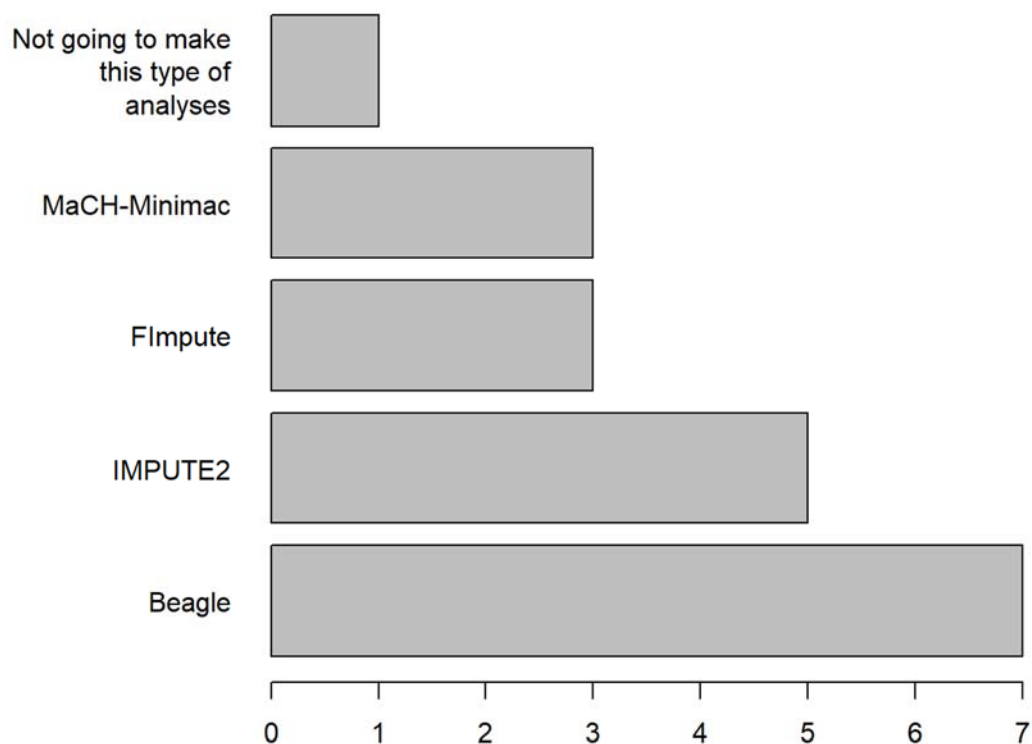


Figure 30 - Usage of tools to perform imputation of phased genotype data. Bars represent the number of times each option was selected by participants.

3.2.9.4 - For association testing which of these tools do you use?

Possible multiple-choice options were: a) PLINK2⁸⁹; b) GenABEL⁹⁰; c) GWASTools⁹¹; d) SNPTTEST⁹⁶; e) QuickTest¹⁰¹; f) GCTA¹⁰²; g) Not going to make this type of analyses; h) Other.

In Figure 31, three tools are the most common choice to perform association testing namely GCTA¹⁰², PLINK2⁸⁹ and GenABEL⁹⁰. Further tools were also reported to be used by a single survey participant.

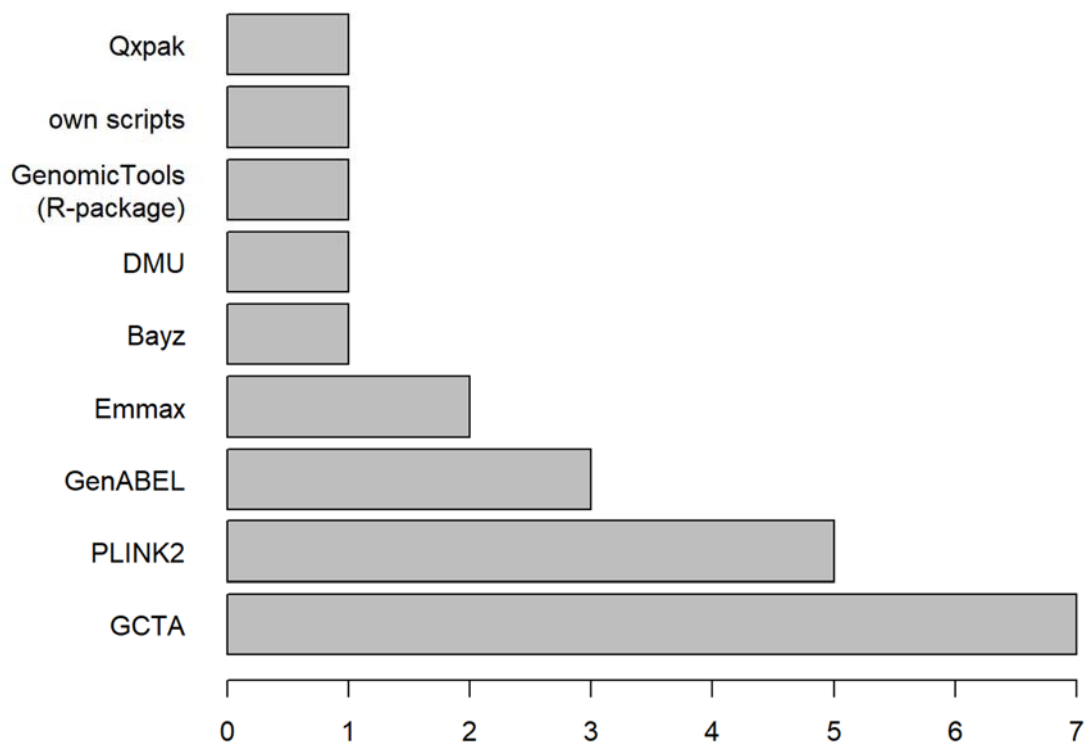


Figure 31- Usage of tools to perform association testing (genotype x phenotype). Bars represent the number of times each option was selected by participants.

4. Conclusions

Section 1

The survey succeeded in receiving responses from 11 BovReg partners in a total of 13 responses.

Section 2

Four BovReg partners do not use **workflow management systems**, and three reported to use BASH, which in comparison with workflow managers as Nextflow or Snakemake assures low reproducibility of outcomes. BovReg partners mostly use open-source software with the exception of **Ingenuity Pathways Analysis (IPA)**. For this case, in WP3, task 3.2, we should consider testing the possibility to include Cytoscape and required modules into the development of reference analysis pipelines after benchmarking its functionality. However, further information is required to be collected from partners in order to identify specifically the types of analysis that are being performed with IPA.

Section 3

Regarding the evaluation of **data quality**, we have observed that most respondents report the use of FASTQC, which, is a widely used tool. However, concerning filtering of data for quality (QC), most of the surveyed partners are using their own scripts. In fact, available open source tools do not allow setting required parameters for RNA-seq or ChIP-Seq data and hence, ad-hoc code had to be developed to this end. This suggests the need for the implementation of a standardized procedure for this data QC analysis step, in BovReg analysis pipelines. Further details should be discussed with the relevant BovReg partners.

Section 4

In this section all partners have reported the use of available open-source tools which are widely used in the community, and do not report the use of own scripts. However, most partners reported the use of more than one read mapper, thus suggesting that for the implementation of BovReg pipelines, selection of read mappers should be discussed with relevant BovReg partners.

Section 6

This section focused on surveying the implementation of methods for the evaluation of **ChIP-Seq quality metrics**. Although available open-source software is being used, many partners interviewed also use their own scripts. This implies that steps such as the estimation of the percentage of reads mapping to peaks or the cross-correlation analysis may require the development of standardized and normalized code for BovReg ChIP-Seq pipelines.

Section 7, 8 & 9

These sections were designed to enquire regarding **specific analyses** related to **ChIP-seq, ATAC-seq and Hi-C data**. Only four partners selected to answer the questions in these sections. Results obtained show that all of these partners have experience in analysing ChIP-Seq data, since they were able to identify tools provided for multiple choice answers, as well as to identify other tools that the survey was not considering.. In contrast, all these partners have responded to lack experience in the analyses of ATAC-seq and Hi-C data types. This indicates that training in these categories of NGS data should be considered for the future BovReg training programme.

Section 10

This section was focused on the assessment of tools being used for the **analysis of RNA-seq data**, and seven of BovReg partners answered to the questions of this section. The respondents were aware of the tools that we have provided for the multiple choice reply options. However, they have reported the use of several tools for performing several of the analysis steps, namely, gene quantification, isoform quantification, normalization and differential expression analysis. Therefore, it is required to further communicate with relevant partners in order for partners involved in WP3, Task3.2 to assess if BovReg pipelines for the analysis of RNA-seq data require such diversity of tools or if we should select one tool for each of these steps.

Section 11

This section was prepared focusing on tools used for the required **analysis workflow of Genome-Wide Association Studies**. Participants recognized and report to use most of the tools available.

Overall Conclusion

In conclusion, we have identified steps of the analysis in this survey that may require the development of normalized and standardized code, however further specific questions are now needed to formulate in order to access which are the required parameters. We have further identified training needs that might be considered by the consortium, in the fields of

- 1) workflow managers and **environment management systems** ;
- 2) ChIP-Seq, ATAC-Seq and Hi-C analysis.

As planned, this survey has enabled us to assess different tools in use by different partners involved. With the replies, we have realized that for some analyses steps partners are using a wide range of tools. It further allowed identifying the need for the development of normalized functions for analyses steps in which partners are using mostly their own

scripts. Next steps will imply to understand the motifs associated with the use of multiple tools for the same analyses as well as to inquire regarding the need to use ad-hoc code for some steps, thus enabling the future development of BovReg pipelines in T3.2.

Furthermore, because of the-clustering activities regarding the bioinformatics tasks of the three FAANG projects funded under the H2020 SFS-30 Agri-Aqua labs call, some needs have changed the original aims of this survey. The leader of task 3.1 and of this deliverable, FMV-ULisboa will additionally make an attempt to elicit input from partners of the other projects in the first cluster workshop taking place February in Hinxton, UK coinciding with the submission deadline of this deliverable. Therefore, the survey will be updated during the runtime of the project.

5. References

1. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
2. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
3. Severin, J. *et al.* eHive: An Artificial Intelligence workflow system for genomic analysis. *BMC Bioinformatics* **11**, (2010).
4. Anaconda, I. Conda. Available at: <https://conda.io/en/latest/>.
5. Docker. Available at: <https://www.docker.com>.
6. Singularity. Available at: <https://singularity.lbl.gov>.
7. Github. Available at: <https://github.com>.
8. GitLab. Available at: https://gitlab.nps.edu/users/sign_in.
9. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. (2010). Available at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. (Accessed: 1st December 2013)
10. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
11. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
12. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

13. Dodt, M., Roehr, J., Ahmed, R. & Dieterich, C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel)*. **1**, 895–905 (2012).
14. Gaspar, J. M. NGMerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 1–9 (2018).
15. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
16. Picard. <http://broadinstitute.github.io/picard/>
17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
19. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics* **CHAPTER**, Unit-11.7 (2010).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
21. Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 283–334 (Springer New York, 2016). doi:10.1007/978-1-4939-3578-9_15
22. Lingala, S. M. & Ghany, M. G. M. Mhs. HISAT: a fast spliced aligner with low memory requirements Daehwan HHS Public Access. *Nat. Methods* **12**, 357–360 (2015).
23. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
24. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
25. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, 1–14 (2010).
26. Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518–2528 (2011).
27. Li, R. *et al.* SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
28. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).

29. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
30. Qin, Q. *et al.* ChiLin: A comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics* **17**, 1–13 (2016).
31. Stark, R. & Brown, G. DiffBind: Differential Binding Analysis. (2011).
32. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
33. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
34. Diaz, A., Nellore, A. & Song, J. S. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol.* **13**, R98 (2012).
35. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–31 (2012).
36. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
37. Genrich. Available at: <https://github.com/jsh58/Genrich>.
38. Gaspar, J. M. Improved peak-calling with MACS2. *bioRxiv* 496521 (2018). doi:10.1101/496521
39. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
40. Wang, C., Xu, J., Zhang, D., Wilson, Z. A. & Zhang, D. An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. *BMC Bioinformatics* **11**, (2010).
41. Kharchenko, P. V, Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
42. Muiño, J. M., Kaufmann, K., van Ham, R. C. H. J., Angenent, G. C. & Krajewski, P. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant Methods* **7**, 11 (2011).
43. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
44. Xu, H. *et al.* A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **26**, 1199–1204 (2010).
45. Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. & Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of

- enrichment, even within amplified genomic regions. *Genome Biol.* **12**, R67 (2011).
46. Song, Q. & Smith, A. D. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* **27**, 870–871 (2011).
 47. Guttman, M. *et al.* Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
 48. Wilder, S. P. SWEMBL.
 49. Rozowsky, J. *et al.* PeakSeq: Systematic Scoring of ChIP-Seq Experiments Relative to Controls. *Nat Biotechnol.* **27**, 66–75 (2009).
 50. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300 (2008).
 51. Nix, D. A., Courdy, S. J. & Boucher, K. M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
 52. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
 53. Mendoza-Parra, M. A., Sankar, M., Walia, M. & Gronemeyer, H. POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Res.* **40**, (2012).
 54. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
 55. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
 56. Liang & K. DBChIP: Differential Binding of Transcription Factor with ChIP-seq.R package version 1.30.0. (2019).
 57. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* **13**, R16 (2012).
 58. Taslim, C., Huang, T. & Lin, S. DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. *Bioinformatics* **27**, 1569–1570 (2011).
 59. Xu, H. & Sung, W. Identifying Differential Histone Modification Sites from ChIP-seq Data. in *Methods in Molecular Biology (Methods and Protocols)* (2012). doi:https://doi.org/10.1007/978-1-61779-400-1_19
 60. Shin, H., Liu, T., Manrai, A. K. & Liu, X. S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605–2606 (2009).

61. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).
62. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
63. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
64. Anders, S. Htseq: Analysing high-throughput sequencing data with python. (2010). Available at: <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>.
65. Fonseca, N., Petryszak, R., Marioni, J. & Brazma, A. iRAP - an integrated RNA-seq Analysis Pipeline. (2014). doi:10.1101/005991
66. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
67. Forster, S. C., Finkel, A. M., Gould, J. A. & Hertzog, P. J. RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics* **29**, 810–812 (2013).
68. Nariai, N. *et al.* TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. *BMC Genomics* **15**, S5 (2014).
69. Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721–1728 (2012).
70. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
71. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32**, i192–i200 (2016).
72. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
73. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–9 (2011).
74. Patro, R., Mount, S. M. & Kingsford, C. Seq Reads Using Lightweight Algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
75. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
76. Lee, S. *et al.* Accurate quantification of transcriptome from RNA-Seq data by

- effective length normalization. *Nucleic Acids Res.* **39**, e9–e9 (2010).
77. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25
 78. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8**, 1765–86 (2013).
 79. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
 80. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–23 (2011).
 81. Wucher, V. *et al.* FEELnc: A tool for Long non-coding RNAs annotation and its application to the dog transcriptome. *bioRxiv* 64436 (2016). doi:10.1101/064436
 82. Sun, L., Liu, H., Zhang, L. & Meng, J. IncRScan-SVM: A tool for predicting long non-coding RNAs using support vector machine. *PLoS One* **10**, 1–16 (2015).
 83. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41**, e166–e166 (2013).
 84. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
 85. Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15**, 311 (2014).
 86. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
 87. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
 88. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
 89. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
 90. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
 91. Gogarten, S. M. *et al.* GWASTools. *Bioinformatics* **28**, 3329–3331 (2012).
 92. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
 93. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from

- Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
94. Sharp, K., Kretzschmar, W., Delaneau, O. & Marchini, J. Phasing for medical sequencing using rare variants and large haplotype reference panels. *Bioinformatics* **32**, 1974–1980 (2016).
 95. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
 96. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
 97. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
 98. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *bioRxiv* 797944 (2019). doi:10.1101/797944
 99. Clayton, D. SNPMatrix. Available at: <http://www.bioconductor.org/packages//2.7/bioc/html/snpMatrix.html>.
 100. Clayton, D. SNPstats.
 101. Kutalik, Z. *et al.* Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics* **12**, 1–17 (2010).
 102. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

6. Annexes

List of participants in the survey:

Name	Institution
Aniek Bouwman	WUR
Andreia Amaral	FMV-ULisboa
Christa Kühn	FBN
Daniel Fischer	LUKE
Emily Clark	UEDIN
Garth Isley	EMBL-EBI
Gabriel Costa	ULIEGE
Goutam Sahana	AU
Graham Plastow	UAL
Hubert Pausch	ETH
José Espinosa	CRG
Mogens Lund	AU.
Romain Philippe	INRAE