

# A pipeline for analysis of allele specific expression from RNA-seq data reveals salinity-dependent response in Nile tilapia

Aurora Campo<sup>1,\*</sup>, Moran Gershoni<sup>1</sup>, Adi Doron-Faigenboim<sup>2</sup>, Avner Cnaani<sup>1,\*</sup>

1. Institute of Animal Science, Agricultural Research Organization, Rishon LeZion, Israel

2. Institute of Plant Science, Agricultural Research Organization, Rishon LeZion, Israel

## Abstract

Species living in a changing environment are capable of adapting to alterations of various factors. Physiological acclimatization may be significantly influenced by the heterozygosity, especially with regards to allele variance and its specific expression (ASE) under different conditions. Data from RNA-seq experiments can be used to identify and quantify the alleles expressed, in order to detect and characterize ASE and regulation of gene expression. However, the allele matching the reference genome creates a mapping bias that prevents a reliable estimation of the allele depth unless the haplotype of the experimental individuals is provided. We developed a pipeline that allows the identification of the alleles corresponding to an RNA-seq dataset and their unbiased quantification. This pipeline does not require the sequencing of the DNA nor the previous knowledge of the haplotype. The identified SNPs are further substituted in the reference genome, thus creating two pseudogenomes with the alternative alleles on two independent samples of the experiment. The SNPs are further called against each pseudogenome thus providing with two SNP datasets that are averaged for calculation of the allele depth. The final SNP calling file contains the coordinates of the SNPs and also the ID of genes containing the SNPs, the expressed genotypes, the unbiased allele depth and the statistical tests for identifying ASE according to the experimental design and correlated with differentially expressed genes. Therefore, the pipeline presented here can calculate ASE in non-model organisms and can be applied to previous RNA-seq datasets for expanding studies in gene expression regulation.

## Introduction

High-throughput RNA-seq is a common technique in many researches, providing differential gene expression (DEGs) data for particular conditions or experimental factors (Marioni *et al.*, 2008). The quantification of gene expression for each factor is based on the counts of the reads that correspond to a particular gene. The sequence of those reads include the variants expressed under the different

34 experimental factors, and therefore it is possible to quantify them (Garber *et al.*, 2011; Oszolak and  
35 Milos, 2011; Trapnell *et al.*, 2011). This allele expression related to a particular factor is known as  
36 Allele Specific Expression (ASE). ASE in a particular condition is one of the ways in which the  
37 organism can respond to the changing environment. This ability is attributed to the individual's  
38 heterozygosity and emphasize the importance of genetic variation as a mechanism of adaptation  
39 (Lande and Shannon, 1996; Hermisson and Pennings, 2005; Barrett and Schluter, 2008; Bernatchez,  
40 2016).

41 There are few studies on the effect of ASE-SNPs under different environmental conditions. These  
42 were mostly related to specific regions of regulation affected by SNPs, also known as expression  
43 quantitative trait loci or eQTLs (Wang, 2017; Zhang *et al.*, 2020). Interestingly, Knowles *et al.* (2017)  
44 developed a generalized linear model tool for analysing genome x environment interactions for ASE,  
45 known as EAGLE. However, this approach is designed for quantitative factors in human or model  
46 organisms, where many genomic tools and genotype datasets are widely available.

47 Another determinant factor for ASE is the tissue, as shown in cattle (Chamberlain *et al.*, 2015). Also  
48 in cattle, Guillocheau *et al.* <sup>(2019)</sup> found that 13% of the total expressed genes in muscle had SNPs in  
49 ASE associated with phenotypic traits and potentially causative of *cis*-regulation. In teleosts, SNP  
50 studies discovered the sex determination patterns of ASE in turbot (*Scophthalmus maximus*)  
51 (Martínez *et al.*, 2019), SNP markers in Atlantic salmon (*Salmo salar*) with higher performance for  
52 DHA (Horn *et al.*, 2020), eQTL affecting resistance to lice in Atlantic salmon (Robledo *et al.*, 2019)  
53 and detection of broad scale suppression of gene expression in triploid medaka (*Oryzias latipes*)  
54 (Garcia *et al.*, 2014). Therefore, ASE is a good estimator of tissues modifications under  
55 environmental factors.

56 A reference genome is used to identify the chromosome and position of the RNA-seq reads after the  
57 alignment of reads and genome sequences. This procedure is known as mapping. The most common  
58 challenge of this approach appears when mapping two different alleles, from which, one is identical  
59 to the same reference sequence. The alignment mismatch between the non-identical allele mapped  
60 against the genome will discard some of the alternative alleles. Therefore, there is a bias towards the  
61 identical or reference allele since some of the reads including the alternative allele are discarded  
62 (Degner *et al.*, 2009). Due to this mapping bias, it is difficult to find the regulatory effects of ASE-  
63 SNPs in gene expression experiments (Monsu and Comin, 2021; Zhan, Griswold and Lukens, 2021).  
64 Solving the mapping bias issue requires the knowledge of the sample haplotype, either from DNA-  
65 sequencing or by using available genotype data and reference haplotypes, such as HapMap  
66 (Consortium, 2003) and other SNP panels (Rozowsky *et al.*, 2011; Vijaya Satya, Zavaljevski and  
67 Reifman, 2012). Some new approaches indicate the utilization of many reference genomes in order  
68 to provide with a broader view of the SNPs in the population (Chen *et al.*, 2021). Unfortunately, these

approaches can't apply to RNA-seq experiments designed without considering sampling and sequencing genomic DNA for haplotype discovery.

We developed a pipeline for SNPs calling and analysis of ASE, using RNA-seq datasets retrieved in experiments aiming to characterize DEGs under different environmental conditions. This pipeline enables quantification of ASE in sampled organisms for which there is no prior genotypic knowledge. We solved the mapping bias without accessing the haplotype of the sampled animals and provide the distribution of alleles in ASE. Our approach creates two pseudogenomes based on allele variants of two samples from different experimental groups. The retrieved SNP dataset can be then submitted to statistical tests for association of allelic expression and environmental or physiological factor. Finally, it is possible to correlate the coordinates of the ASE SNPs with other data on the gene expression such as DEGs and metilome sites to complement the results.

In the present article we apply our pipeline to study the effect of high salinity challenge on a freshwater fish, the Nile tilapia (*Oreochromis niloticus*). We performed a discovery and unbiased quantification of bi-allelic sites and statistical assessment of SNPs in ASE in two tissues, gills and kidney, and two environmental factors, freshwater and brackish water.

## Material and methods

### *Ethical statement*

This study was approved by the Agricultural Research Organization Committee for Ethics in Experimental Animal Use, and was carried out in compliance with the current laws governing biological research in Israel (Approval number: IL-715/17).

### *Samples origin, processing and sequencing*

The sequences used in this study were from an experiment previously described by Root et al. (2021a, 2021b). Briefly, Twelve Nile tilapia male fish were randomly distributed between two 600 l freshwater tanks. After 2-week acclimation, one group was exposed to a gradual salinity increase of 5 ppt per day up to a final salinity of 25 ppt. Gills and kidney samples were taken after 24 h at the final salinity. mRNA was extracted using TRIzol reagent (Thermo Fisher Scientific), and purified to remove DNA contamination using the TURBO DNA-free kit (Invitrogen). Total mRNA samples were sent to the Israel National Center for Personalized Medicine (INCPM) at the Weizmann Institute of Science (Rehovot, Israel), where quality was determined on TapeStation Agilent 2200 system, before library preparation and sequencing on an Illumina Hi-Seq 2500 device.

For validation of the SNP calling, 8 tilapia individuals were sampled. RNA was extracted from the gills and for genomic DNA was extracted from fin clips, using RNeasy mini kit and DNeasy blood and tissue kit (Qiagen, Hilden, Germany), respectively. Sequencing of RNA and DNA was performed

104 with Illumina NovaSeq S1 300 including UMI barcoding at 10x and 30x coverage respectively, in  
105 the INCPM.

106

#### 107 *Pipeline for mapping bias removal by the use of pseudogenomes and SNP calling*

108 A pipeline developed in Snakemake (Köster and Rahmann, 2012) is proposed here for retrieving  
109 SNPs from transcriptome analysis, while eliminating the mapping bias without previous knowledge  
110 of the genotypes (Figure 1, Supplementary 1). The entire code with scripts to the pipeline is available  
111 at GitHub ([https://github.com/AylaScientist/Snakemake\\_for\\_SNPs](https://github.com/AylaScientist/Snakemake_for_SNPs)). *Fastq* files received from the  
112 INCPM were processed according to the proposed pipeline: The *fastq* files were trimmed with  
113 Trimmomatic (Bolger, Lohse and Usadel, 2014) and quality was verified with FASTQC (v0.11.8,  
114 Andrews, 2010). The trimmed *fastq* files mapped against the reference Genome of *O. niloticus*  
115 (NMBU GCF\_001858045.2) with the RNA-seq alignment tool STAR (v2.7.1a, Dobin et al., 2013).  
116 SNPs were called following GATK best practices (Poplin *et al.*, 2017) as described in GATK best  
117 practices (<https://github.com/gatk-workflows/gatk4-rnaseq-germline-snps-indels>). Two samples  
118 from different experimental groups were chosen for construction of two pseudogenomes from a *vcf*  
119 file, following the protocol by Johan Zicola ([https://github.com/johanzi/make\\_pseudogenome](https://github.com/johanzi/make_pseudogenome), MIT  
120 license). *Fastq* files were processed against the pseudogenomes described above. Two final *vcf* files  
121 joining the SNPs from all samples were annotated using ANNOVAR (Wang, Li and Hakonarson,  
122 2010). In order to annotate the SNPs of the non-model species Nile tilapia, we constructed a database  
123 using the annotation file from the same release as the genome of reference and the ANNOVAR scripts  
124 meant for creating such a database. Allele depth and genotype were collected into a table  
125 (VariantsToTable, GATK). The two datasets were then submitted to home developed scripts in  
126 Python v3.7.3. These scripts calculate the average counts for the reference and alternative allele,  
127 calculate the allele frequency, checks the correction of the mapping bias and develop the statistical  
128 analysis according to the experimental design. The home made scripts are part of the pipeline  
129 designed in snakemake and can be found in the release on gitHub.

130

#### 131 *Data engineering and statistical analysis*

132 The pipeline provided with two datasets, each one containing the SNP sites called to one of the  
133 pseudogenomes. The two datasets were merged and filtered for multiallelic sites with Python using  
134 pandas and Numpy specific for data science methods that can be found in the gitHub. Afterwards, the  
135 counts of each reference and alternative polymorphic site were averaged (Figure 2A). SNPs for which  
136 the depth of one allele was less than 3 and SNPs for which the total allele depth from reference and  
137 alternative alleles was less than 10, were deleted. Those sites that had a monoallelic expression were  
138 removed. Only SNPs shared by all individuals were left on the final data of consensus SNPs.

139 These resultant SNPs were submitted to statistical analysis for Allele Specific Expression of treatment  
140 using a Chi-square test for comparison between salinity treatments. Each experimental group was  
141 compared with each other leading to four Chi-square tests (Table 1). The p-values were adjusted with  
142 the Bonferroni test using the python library Multitest.

143 Significant ASE SNPs for each treatment were analysed for GO function of its gene. DEG analysis  
144 was performed with DESeq package (Anders *et al.*, 2010) in R (v 3.6.3, **Development Core Team,**  
145 **2013**) for salinity. In order to find regulatory pathways, the SNPs in ASE where contrasted with the  
146 significant DEGs.

147

148 *Validation by sequencing / re-sequencing:*

149 For validation of SNP calling by the above described pipeline, eight additional fish were sampled.  
150 RNA extracted from the gills and DNA extracted from fins of each individual were sequenced and  
151 processed with the pipeline for SNP calling as described before in the text. The retrieved SNPs from  
152 DNA were selected for exonic single nucleotide variant (SNVs), thus avoiding indels and intronic  
153 sites. The monoallelic expression was also deleted in order to obtain normal distributed data for the  
154 allele frequency. The selection of SNVs obtained with the pseudogenomes was contrasted with the  
155 retrieved SNVs from RNA obtained with the reference using a T-Student test. A deeper analysis with  
156 IGV (Robinson *et al.*, 2017) was performed with 20 SNV among false positives.

157

## 158 **Results**

159

160 We were able to determine 103,843 informative SNPs from our experimental population present in  
161 all the tilapia individuals, from which, 99,885 present monoallelic expression and 3,740 follow a  
162 normal distribution.

163 The method for SNP calling was tested by comparing abundance distribution of the allele frequencies.  
164 The comparison was performed on the SNPs that didn't present monoallelic expression. There are  
165 two clearly different distributions ( $p < 0.01$ ), one for classical SNP calling against reference genome  
166 and the other for the calling against two pseudogenomes developed in this study (Figure 3, table 2).  
167 The SNPs called in the kidney show a normal distribution of the frequency for both calling methods  
168 (Supplementary 5, F and H). In these groups the most frequent alleles called after the reference  
169 genome are at 0.65 and 0.6 in the fresh and salty water respectively. In the gills, the calling on the  
170 reference genome produced three different distributions. One is biased towards the reference allele,  
171 the second is biased towards the alternative allele and a third is a normal distribution. The highest  
172 frequency of the alleles was biased mostly towards the reference allele but also to the alternative. The  
173 alleles that show a normal distribution include the smallest number of alleles. The average allele

174 frequency in the normal distribution shows of 0.55 in the fresh and salty water groups (Supplementary  
175 5, B and D). The SNP called after mapping to the pseudogenomes shows unbiased normal distribution  
176 marked by the highest frequency of the alleles at 0.5 for all the studied experimental groups. The  
177 Student T-test shows significant differences in the distributions of the allele frequencies (Figure 3,  
178 table 2,  $p < 0.01$ ).

179 Chi-square tests indicated ASE for the different salinities tested in the gills and the kidney on the  
180 SNPs that do not show monoallelic expression (Table 1, Supplementary 1, 2, 3 and 4). The significant  
181 ASE SNPs were classified according to their function (Figure 4, table 3). Nearly all SNPs were from  
182 non-coding regions correspond to 3'UTR, non-coding region and 5'UTR for all the tests.  
183 Substitutions occur in synonymous and non-synonymous variants with higher frequency for  
184 synonymous. Few upstream/downstream, stop loss or stop gain, frameshift or non-frameshift  
185 insertion or deletion were found. Additionally, some variants are found to be assigned to intronic,  
186 intergenic upstream and downstream variants (Figure 4, table 3).

187 The analysis of differentially expressed genes indicated 899 SNPs corresponding to also differentially  
188 expressed genes when comparing gills and kidney in fresh water (test 1), and 1,153 SNPs in  
189 differentially expressed genes between gills and kidney in salty water (test 2) (Table 1, Supplementary  
190 6 and 7 respectively). From these, 629 (69.6%) and 790 (68.5%) correspond to regulatory regions  
191 such as UTRs and ncRNAs for tests 1 and 2 respectively. When comparing the salinity, there were  
192 15 DEG with ASE SNPs in the gills (test 3) and none in the kidney (test 4). From these, 12 SNPs  
193 (80%) are located in 3'UTR as the only regulatory region (Supplementary 8). No ASE SNPs with  
194 differentially expressed genes in the kidney have been retrieved.

195 The CHI-tests (Figure 5) show 50 common SNPs in ASE independent of tissue or salinity treatment.  
196 There are 929 SNPs in ASE found differentially expressed between gills and kidney, independently  
197 of the salinity conditions. No SNPs in ASE were found common uniquely to the effect of the salinity  
198 in gills and in kidney.

199 The comparative heatmap of the allele frequencies shows a differential pattern for tissues and salinity  
200 (Figure 6).

201 The function described by Gene Ontology (GO) analysis in the ASE SNPs was compared for the  
202 salinity challenge and the tissue differences. The ASE variants between tissues conserve a similar  
203 proportion of functions both in fresh and salty water. On the other hand, the ASE SNPs between  
204 salinities change the gene function within the kidney and within the gills (Figure 7).

205 The chromosomic regions of interest for significant ASE SNPs are illustrated in the Manhattan plot  
206 (Figure 8).

207

208 *Validation by sequencing*

209 The SNP calling through the pipeline from the gills transcriptome sequences of 8 Nile tilapia  
210 individuals resulted in 85% of them also called from the genome sequences. We performed an  
211 analysis of 20 single nucleotide variants (SNV) from the 15% of the SNPs not corresponding to the  
212 genome, by visualization with IGV software (Robinson *et al.*, 2017). SNVs called from the  
213 transcriptome showed to be false SNPs in 14 cases, from which 7 cases where the allele counts were  
214 below 5. Additionally, 6 SNVs proved to exist in the RNA. From those, 1 SNV was also present in  
215 the DNA, and the other 5 SNVs were present only in RNA.

216

217

## 218 **Data availability**

219 The sequencing data was submitted to SRA under the bioproject PRJNA669315. The snakemake  
220 pipeline is submitted to the GitHub [https://github.com/AylaScientist/snakemake\\_for\\_SNPs](https://github.com/AylaScientist/snakemake_for_SNPs)

221

## 222 **Discussion**

223 Two methods for SNP calling were compared in the present study. The first method is the commonly  
224 used, which includes the mapping of the reads to the reference genome previous to the SNP calling.  
225 As our results show, the allele frequency of the kidney follows a normal distribution with a slight bias  
226 towards the reference genome marked by the highest density of allele frequency at 60% reference  
227 allele versus 40% alternative allele (Supplementary 5 F and H). In the gills, most of the alleles follow  
228 a binomial distribution, including some monoallelic expression and bias towards both the reference  
229 and the alternative alleles (Supplementary 5 B and D). In the second method, using our new approach,  
230 the SNP calling takes place after mapping the reads to pseudogenomes (Figure 1). These  
231 pseudogenomes contain the SNPs expressed in the experimental set. Our result shows that the allele  
232 frequency of the average counts on these SNPs will follow an unbiased normal distribution (Figure  
233 3, supplementary 5, table 2,  $p < 0.01$ ).

234 Previous strategies for removing mapping bias require prior knowledge of genotypes (Rozowsky *et al.*  
235 *et al.*, 2011; Yuan and Qin, 2012; Pandey *et al.*, 2013; Xin *et al.*, 2013; Mayba *et al.*, 2014; Braasch *et al.*,  
236 *et al.*, 2016; Guillocheau *et al.*, 2019), elimination of sites showing bias after simulation (Pickrell *et al.*,  
237 2010; Stevenson, Coolon and Wittkopp, 2013; Panousis *et al.*, 2014; Hodgkinson *et al.*, 2016), the  
238 SNPs previously informed in a panel (Van De Geijn *et al.*, 2015; Salavati *et al.*, 2019; Gutierrez-  
239 Arcelus *et al.*, 2020) or direct use of a variant-aware alignment (Hach *et al.*, 2014; Buchkovich *et al.*,  
240 2015; Miao *et al.*, 2018). The pipeline developed in this study does not require this previous  
241 knowledge. Instead, it detects the sites expressed in at least one of the individuals in the experiment.  
242 This detection takes place after mapping to the reference genome previous to SNP calling (Figure 1).

243 The sites revealed in this first calling may correspond to alleles found in the genotype, but also to few  
244 SNPs generated after mRNA modifications or editing.

245 Editing of RNA consists of discrete changes to specific nucleotide sequences within an RNA molecule  
246 after it has been transcribed by RNA polymerase (Maas and Rich, 2000; Kiss, 2001). This molecular  
247 process is found in every living organism and it is evolutionary conserved (Song, Yi and He, 2012;  
248 Li and Mason, 2014; Meyer and Jaffrey, 2014; Sun *et al.*, 2016). It can include deamination of single  
249 sites leading the substitution of cytidine (C) to uridine (U) and adenosine (A) to inosine (Takenaka *et*  
250 *al.*, 2014; Shikanai, 2015; Licht *et al.*, 2016; Licht, Hartl, *et al.*, 2019; Licht, Kapoor, *et al.*, 2019) but  
251 also generalized insertions and deletions of uracil in the same transcript by an editosome, also known  
252 as pan-editing (Blum, Bakalara and Simpson, 1990; Stuart, 1991; Benne, 1994; Simpson and  
253 Thiemann, 1995; Jan Arts and Benne, 1996; Alfonzo, Thiemann and Simpson, 1997; Kable,  
254 Heidmann and Stuart, 1997). If the editing takes place in the mRNA it can derive in the modification  
255 of the aminoacid sequence of the protein encoded (Brennicke, Marchfelder and Binder, 1999).

256 These editions in RNA can modify the cell biology by modifying the RNA structure, tuning  
257 interactions within the ribosome and by recruiting specific binding proteins that intersect with other  
258 signalling pathways (Nachtergaele and He, 2017). Interestingly, they are also dynamic, changing in  
259 distribution or level in response to stresses, such as heat shock and nutrient deprivation (Carlile *et al.*,  
260 2014; Schwartz *et al.*, 2014; Li *et al.*, 2015), translation control in immune processes (Piccirillo *et al.*,  
261 2014; Araki *et al.*, 2017; Wolf *et al.*, 2020), during cancer proliferation (Gingold *et al.*, 2014; Zviran  
262 *et al.*, 2019), post-transcriptional modifications in development and stem cells (Frye and Blanco,  
263 2016) and during physiologically normal proliferation of T cells (Rak *et al.*, 2021). In our pipeline,  
264 the SNPs found in the mRNA belong to the expression under experimental conditions. These variants  
265 can include both genomic alleles and post-transcriptional editions that are substituted into the  
266 reference genome thus creating a pseudogenome.

267 Finally, the constitution of two pseudogenomes with RNA expressed under two different conditions  
268 of the study compiles a wider scope of the variability in the analysis. By mapping to the  
269 pseudogenomes, the pipeline developed here may allow the unbiased quantification of the SNPs in  
270 the genotype and of the post-transcriptional modifications of the mRNA also. We performed a  
271 validation of the SNV sites identified by our pipeline towards the genotype of a control population of  
272 tilapia exposed to fresh water. The results indicate that 85% of the SNVs are correctly called after an  
273 existing genotype. The study of 20 SNV sites among the 15% that were not found in the genotype  
274 revealed that only 60% of them are false positives and mostly related to a low count of the reads.  
275 Interestingly, 30% of these SNVs were consistently expressed in the sequenced mRNA and the allele  
276 depth estimation allowed a correct allelic imbalance estimated by the pipeline. These results indicate  
277 a possible mRNA editing among the sources of false positives. Consequently, the SNPs in ASE from



our analysis may include the variant sites whose expression and modification is regulated under the salinity challenge and the different studied tissues, gills and kidney.

Our analysis on the allele frequency indicates two types of imbalanced SNPs: monoallelic expression and normal distribution (Supplementary 9 A). The monoallelic expression is represented by the allele frequencies 0 or 1 in heterozygote sites indicating allele imprinting. Tissue-specific imprinting was described before in human and mouse (Babak *et al.*, 2015) concluding that nearly all the imprinted alleles were imprinted in early development. In our analysis, when MAE alleles detected in one tissue, they show higher allelic imbalance towards the reference allele in another studied tissue (Supplementary 9 A). Such distribution has been described previously and was called variable ASE (Skelly *et al.*, 2011). Variable ASE is represented by non-normal distribution of the allele frequencies in the graphs (Supplementary 9 A). This distribution is consistent with the one described by Skelly *et al.* (2011), indicating greater dispersion in read counts after differential exon expression. This indicates complex patterns of ASE, such as allele specific alternative splicing. Tissue-specific genetic control of splicing have been described in humans for polymorphisms affecting splicing and expression in human blood and brain tissues (Heinzen *et al.*, 2008). Tissue-specific isomorphs can be regulated by alternative polyadenylation of the 3'UTR length in human (Weng *et al.*, 2016; Macdonald, 2019), *Drosophila* (Sanfilippo, Wen and Lai, 2017), *Caenorhabditis elegans* (Khraiwesh and Salehi-ashtiani, 2017) and yeast (Liu *et al.*, 2017). Our analysis indicates that a tissue-specific imprinting and splicing may occur in gills and kidney of Nile tilapia driven by ASE. Interestingly, when MAE sites are strictly filtered (Supplementary 9 B) the variable ASE is also filtered, meaning that the SNPs in charge of the variable ASE may be also related to the monoallelic expression in other tissues. Further analyses on imprinted genes may illustrate this phenomenon and evaluate the network of MAE genes associated to variable ASE phenomenon, especially in regard to the tissue function under environmental challenge.

### *Salinity challenge in tilapia*

In the present study we obtained an unbiased counting of allele expressed in different tissues, gills and kidney, after exposure to salinity challenge. The process followed the GATK best practices recommendations (Poplin *et al.*, 2017) and provided with unbiased SNPs from which 236 are associated to the salinity challenge in gills and 1,126 in the kidney. Other algorithms depending on DNaseq data for calculating ASE have also established genomic x environmental interactions, as for example the EAGLE tool (Knowles *et al.*, 2017). This tool is only applying to certain model organisms and it provided with 442 ASE SNPs (associations in the article) for the reaction of the human liver to different molecules. Therefore, the number of ASE SNPs retrieved after the

environmental challenge in our pipeline with non-model organism are within the range of results obtained with tools limited to model organisms.

The statistical analysis showed more SNPs in ASE for the tests between tissues than for the tests on the salinity challenge, independently of the salinity process. The GO functions of the genes containing the ASE SNPs between gills and kidney are highly similar between both salinities tested, also in proportion of functions. Previous studies of cattle ASE SNPs, in tissues of one individual, evaluated the allelic imbalance within each tissue. This analysis reported that at minimum 89% of the total SNPs were imbalanced in at least one tissue out of 18 studied (Chamberlain *et al.*, 2015). Allelic imbalance was also common between 19 muscles samples of the Limousine cattle breed (Guillocheau *et al.*, 2019). Tissue-specific regulation of allele expression was also studied in mouse allelome (Andergassen *et al.*, 2017), finding that the regulation of ASE may be driven by tissue-specific enhancers or by post-transcriptional differences. In our study we also find a basal regulation of this tissue-specific allelic expression affecting 1,589 SNPs for gills and kidney independently of the salinity (Figure 5 and Figure 6). More epigenomic analyses are needed for testing if there is tissue specific epigenomic regulation of SNP expression such imprinting in tilapia, as previously suggested for mouse development (Andergassen *et al.*, 2017).

The number of SNPs in ASE for the challenged gills (236 SNPs) is about a quarter of the SNPs in ASE for the kidney (1,126 SNPs) (Table 1). Some of the SNPs in ASE were associated with differentially expressed genes. Both gills and kidney have SNPs in ASE related to protein binding, membrane and integral components of membrane, membrane and oxidation-reduction process (Figure 7). On the other hand, gills change the expression of SNPs in genes associated to tricarboxylic acid cycle, transmembrane transport and oxoglutarate dehydrogenase activity that is not present in the kidney. Previous transcriptomic and proteomic analysis on these data indicated that there is a response in the gills to salinity by differential expression of genes related to epithelium turnover (Root, Campo, Macniven, Con, Cnaani and Kùltz, 2021; Campo *et al.*, 2022). Not only that, the proteomic analysis revealed higher post-translational modifications in the kidney as a response to the salinity exposure in contrast with few differentially expressed genes (Root, Campo, Macniven, Con, Cnaani and Kùltz, 2021). These results are consistent with this complementary analysis where ASE SNPs are associated to DEGs in the gills but not in the kidney. All taken together may indicate the differential expression found in gills to cope with the salinity challenge may be regulated partially by ASE SNP, thus driving the epithelium turnover. Our results suggest that the salinity, as environmental factor, may challenge each tissue in a different manner. While the response in gills correspond to a higher DEG, the response in the kidney provides with higher number of ASE SNPs.

The 3'UTR SNPs is the most frequent type of SNP found (Figure 4, table 3). The role of 3'UTR in regulation of mRNA was reviewed by Mayr (2017), finding functions of degradation, translation and

347 localization as well as interactions to noncoding and small RNA. Additionally, the functional  
348 interpretation of variants in the 3'UTR has been related to modification of alternative polyadenylation  
349 motifs and RNA-binding protein binding sites, also known as 3'QTLs, and can be used to interpret  
350 16.1% of trait-associated variants in human (Li *et al.*, 2019). Therefore, some of the found SNPs are  
351 likely regulatory ones.

352 After 3'UTR, the second most common SNPs identified for significant allelic expression were found  
353 in codifying regions, mostly synonymous SNPs. The synonymous sites were around 11 to 13 times  
354 more abundant than the non-synonymous in all the tests, except in the comparison of the gills from  
355 fresh to salty water (test 3, table 3), where the ratio of synonymous vs. non-synonymous is ~8.  
356 Diversity among non-synonymous SNPs is significantly lower than among synonymous substitutions  
357 (Graur and Li, 1997) due to the natural selection acting on the non-synonymous SNPs (Ohta, 1995),  
358 and that was the case in all our comparisons. It yet to be determine if different ratios of  
359 synonymous/non-synonymous SNPs in ASE between tissues can indicate different evolutionary  
360 adaptation mechanism between them.

361 Intronic SNPs in ASE were captured in our analysis. Nascent RNAs of longer genes often include  
362 extensive intronic regions that would commonly be removed in the mature RNAs captured in the  
363 whole cells (Mercer *et al.*, 2012; Lake *et al.*, 2017), thus indicating RNA previous to the splicing was  
364 captured. Additionally, the presence of intronic RNAs have been related to transcriptional regulation  
365 events such as splicing and also to cellular identity (Ameur *et al.*, 2011; Gaidatzis *et al.*, 2015; Lake  
366 *et al.*, 2016; Sheng *et al.*, 2017; Yang *et al.*, 2017). Therefore, the significant change in the allele  
367 frequency of intronic SNPs may be related to *de novo* expression of genes and specific splicing  
368 processes depending on tissue and also salinity challenge. The little variation that was found for  
369 upstream/downstream, stop loss or stop gain, frameshift nor non-frameshift insertion or deletion,  
370 indicating that nonsense-mediated decay and other pathological processes are not dominating the  
371 specific expression after salinity exposure on the studied tissues.

372

## 373 **Conclusions**

374 Our pipeline succeeded in providing a robust method on quantification of SNPs that allow the  
375 unbiased determination of SNPs in ASE, under different factors, without the prior knowledge of the  
376 genotype. This approach is suitable for any non-model organism, independently of the strain or the  
377 available genome of reference.

378 Our tool provides with the possibility to reanalyze data of DEGs experiments in order to find gene  
379 regulation and new protein to protein interactions determined by specifically expressed alleles. The  
380 coordinates of the SNPs can be also merged with other sources of transcript data such as methylome.

381 After adaptating of the database for gene annotation, transcriptomes can also be used for SNP calling  
382 in case there is no genome of reference.  
383 In the presented example of use for this pipeline we discovered allelic resources for copying with  
384 salinity exposure in the kidney and in the gills, and that there is differential allelic response to  
385 environment factor, depending on tissue.

386

## 387 **Acknowledgements**

388 This investigation was supported by grant 2016611 from the US-Israel Binational Science  
389 Foundation BSF and the US-Israel Binational Agricultural Research and Development Fund  
390 (BARD) Grant (IS-5358-21).

391

## 392 **Bibliography**

- 393 Alfonzo, J. D., Thiemann, O. and Simpson, L. (1997) 'The mechanism of U insertion / deletion  
394 RNA editing in kinetoplastid mitochondria', *Nucleic Acid Research*, 25(19), pp. 3751–3759.
- 395 Ameer, A. *et al.* (2011) 'Total RNA sequencing reveals nascent transcription and widespread co-  
396 transcriptional splicing in the human brain', *Nature Structural & Molecular Biology*. Nature  
397 Publishing Group, 18(12), pp. 1435–1440. doi: 10.1038/nsmb.2143.
- 398 Andergassen, D. *et al.* (2017) 'Mapping the mouse Allelome reveals tissue-specific regulation of  
399 allelic expression', *eLife*, 6(Xci), pp. 1–29. doi: 10.7554/eLife.25125.
- 400 Anders, S. *et al.* (2010) 'Differential expression analysis for sequence count data via mixtures of  
401 negative binomials', *Advances in Environmental Biology*, 7(10), pp. 2803–2809. Available at:  
402 [http://amsdottorato.unibo.it/6741/1/bonafede\\_elisabetta\\_tesi.pdf](http://amsdottorato.unibo.it/6741/1/bonafede_elisabetta_tesi.pdf).
- 403 Andrews, S. (2010) 'FastQC: a quality control tool for high throughput sequence data'. Babraham  
404 Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- 405 Araki, K. *et al.* (2017) 'Translation is actively regulated during the differentiation of CD8+ effector  
406 T cells', *Nature immunology*, 18(9), pp. 1046–1057. doi: 10.1038/ni.3795.
- 407 Babak, T. *et al.* (2015) 'Genetic conflict reflected in tissue-specific maps of genomic imprinting in  
408 human and mouse', *Nature Genetics*. Nature Publishing Group, 47(5), pp. 544–549. doi:  
409 10.1038/ng.3274.
- 410 Barrett, R. D. H. and Schluter, D. (2008) 'Adaptation from standing genetic variation', *Trends in  
411 Ecology and Evolution*, 23(1), pp. 38–44. doi: 10.1016/j.tree.2007.09.008.
- 412 Benne, R. (1994) 'Review RNA editing in trypanosomes', *European Journal of Biochemistry*, 221,  
413 pp. 9–23.
- 414 Bernatchez, L. (2016) 'On the maintenance of genetic variation and adaptation to environmental

change: considerations from population genomics in fishes', *Journal of Fish Biology*, 89(6), pp. 2519–2556. doi: 10.1111/jfb.13145.

Blum, B., Bakalara, N. and Simpson, L. (1990) 'A model for RNA editing in kinetoplastid mitochondria: RNA molecules transcribed from maxicircle DNA provide the edited information', *Cell*. Elsevier, 60(2), pp. 189–198.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Braasch, I. *et al.* (2016) 'The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons', *Nat Genet*, 48(4), pp. 427–437. doi: 10.1038/ng.3526.

Brennicke, A., Marchfelder, A. and Binder, S. (1999) 'RNA editing', *FEMS Microbiology Reviews*, 23, pp. 297–316.

Buchkovich, M. L. *et al.* (2015) 'Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci', *BMC Medical Genomics*. BMC Medical Genomics, 8(1), pp. 1–15. doi: 10.1186/s12920-015-0117-x.

Campo, A. *et al.* (2022) 'Different transcriptomic architecture of the gill epithelia in Nile and Mozambique tilapia after salinity challenge', *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, 41, p. 100927. doi: <https://doi.org/10.1016/j.cbd.2021.100927>.

Carlile, T. M. *et al.* (2014) 'Pseudoridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells', *Nature*. Nature Publishing Group, 515(7525), pp. 143–146. doi: 10.1038/nature13802.

Chamberlain, A. J. *et al.* (2015) 'Extensive variation between tissues in allele specific expression in an outbred mammal', *BMC Genomics*. BMC Genomics, 16(1), pp. 1–20. doi: 10.1186/s12864-015-2174-0.

Chen, N. *et al.* (2021) 'Reference flow : reducing reference bias using multiple population genomes', *Genome Biology*. Genome Biology, 22(8), pp. 1–17. doi: <https://doi.org/10.1186/s13059-020-02229-3>.

Consortium, I. H. (2003) 'A haplotype map of the human genome The International HapMap Consortium', *Physiological Genomics*, 437(7063), pp. 1299–320. Available at: <http://physiolgenomics.physiology.org/cgi/content/full/13/1/3>.

Degner, J. F. *et al.* (2009) 'Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data', *Bioinformatics*, 25(24), pp. 3207–3212. doi: 10.1093/bioinformatics/btp579.

Development Core Team, R. (2013) 'R: A language and environment for statistical computing'. Vienna, Austria: R Foundation for statistical computing.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp.

15–21. doi: 10.1093/bioinformatics/bts635.

Frye, M. and Blanco, S. (2016) ‘Post-transcriptional modifications in development and stem cells’, *Development*, 3, pp. 3871–3881. doi: 10.1242/dev.136556.

Gaidatzis, D. *et al.* (2015) ‘Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation’, *Nature Biotechnology*. Nature Publishing Group, 33(7), pp. 1–10. doi: 10.1038/nbt.3269.

Garber, M. *et al.* (2011) ‘Computational methods for transcriptome annotation and quantification using RNA-seq’, *Nature Methods*. Nature Publishing Group, 8(6), pp. 469–477. doi: 10.1038/nmeth.1613.

Garcia, T. I. *et al.* (2014) ‘Novel method for analysis of allele specific expression in triploid *Oryzias latipes* reveals consistent pattern of allele exclusion’, *PLoS ONE*, 9(6), p. 100250. doi: 10.1371/journal.pone.0100250.

Van De Geijn, B. *et al.* (2015) ‘WASP: Allele-specific software for robust molecular quantitative trait locus discovery’, *Nature Methods*, 12(11), pp. 1061–1063. doi: 10.1038/nmeth.3582.

Gingold, H. *et al.* (2014) ‘A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation’, *Cell*. Elsevier Inc., 158(6), pp. 1281–1292. doi: 10.1016/j.cell.2014.08.011.

Graur, D. and Li, W. (1997) *Fundamentals of molecular evolution*. second. Sunderland, MA: Sinauer Associates, Incorporated Publishers.

Guillocheau, G. M. *et al.* (2019) ‘Survey of allele specific expression in bovine muscle’, *Scientific Reports*, 9(1), pp. 1–11. doi: 10.1038/s41598-019-40781-6.

Gutierrez-Arcelus, M. *et al.* (2020) ‘Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci’, *Nature Genetics*. Springer US, 52(3), pp. 247–253. doi: 10.1038/s41588-020-0579-4.

Hach, F. *et al.* (2014) ‘MrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications’, *Nucleic Acids Research*, 42(W1), pp. 494–500. doi: 10.1093/nar/gku370.

Heinzen, E. L. *et al.* (2008) ‘Tissue-specific genetic control of splicing: Implications for the study of complex traits’, *PLoS Biology*, 6(12), pp. 2869–2879. doi: 10.1371/journal.pbio.1000001.

Hermisson, J. and Pennings, P. S. (2005) ‘Soft sweeps: Molecular population genetics of adaptation from standing genetic variation’, *Genetics*, 169(4), pp. 2335–2352. doi: 10.1534/genetics.104.036947.

Hodgkinson, A. *et al.* (2016) ‘A haplotype-based normalization technique for the analysis and detection of allele specific expression’, *BMC Bioinformatics*. BMC Bioinformatics, 17(1), pp. 1–10. doi: 10.1186/s12859-016-1238-8.

Horn, S. S. *et al.* (2020) ‘Accuracy of selection for omega-3 fatty acid content in Atlantic salmon fillets’, *Aquaculture*. Elsevier, 519(April 2019), p. 734767. doi: 10.1016/j.aquaculture.2019.734767.

485 Jan Arts, G. and Benne, R. (1996) 'Mechanism and evolution of RNA editing in kinetoplastida',  
486 *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1307(1), pp. 39–54. doi:  
487 [https://doi.org/10.1016/0167-4781\(96\)00021-8](https://doi.org/10.1016/0167-4781(96)00021-8).

488 Kable, M. L., Heidmann, S. and Stuart, K. D. (1997) 'RNA editing: getting U into RNA', *Trends in*  
489 *biochemical sciences*. Elsevier, 22(5), pp. 162–166.

490 Khraiwesh, B. and Salehi-ashtiani, K. (2017) 'Alternative Poly(A) Tails Meet miRNA Targeting in  
491 *Caenorhabditis elegans*', 206(June), pp. 755–756.

492 Kiss, T. (2001) 'Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs',  
493 *The EMBO journal*, 20(14), pp. 3617–3622.

494 Knowles, D. A. *et al.* (2017) 'Allele-specific expression reveals interactions between genetic  
495 variation and environment', *Nature Methods*. Nature Publishing Group, 14(7), pp. 699–702. doi:  
496 10.1038/nmeth.4298.

497 Köster, J. and Rahmann, S. (2012) 'Snakemake-a scalable bioinformatics workflow engine',  
498 *Bioinformatics*, 28(19), pp. 2520–2522. doi: 10.1093/bioinformatics/bts480.

499 Lake, B. B. *et al.* (2016) 'Neuronal subtypes and diversity revealed by single-nucleus RNA  
500 sequencing of the human brain', 352(6293), pp. 352–357.

501 Lake, B. B. *et al.* (2017) 'single-nucleus and single-cell transcriptomes confirms accuracy in  
502 predicted cell-type expression from nuclear RNA', *Scientific Reports*. Springer US, (October 2016),  
503 pp. 1–8. doi: 10.1038/s41598-017-04426-w.

504 Lande, R. and Shannon, S. (1996) 'The Role of Genetic Variation in Adaptation and Population  
505 Persistence in a Changing Environment Author ( s ): Russell Lande and Susan Shannon Published  
506 by : Society for the Study of Evolution Stable URL : <http://www.jstor.org/stable/2410812>  
507 Accessed : 07', *Evolution*, 50(1), pp. 434–437.

508 Li, L. *et al.* (2019) 'Genetic Basis of Alternative Polyadenylation is an Emerging Molecular  
509 Phenotype for Human Traits and Diseases', *bioRxiv*.

510 Li, S. and Mason, C. E. (2014) 'The Pivotal Regulatory Landscape of RNA Modifications', *Annual*  
511 *Review of Genomics and Human Genetics*, 15, pp. 127–150. doi: 10.1146/annurev-genom-090413-  
512 025405.

513 Li, X. *et al.* (2015) 'Chemical pulldown reveals dynamic pseudouridylation of the mammalian  
514 transcriptome', *Nature Chemical Biology*, 11(8), pp. 592–597. doi: 10.1038/nchembio.1836.

515 Licht, K. *et al.* (2016) 'Adenosine to Inosine editing frequency controlled by splicing efficiency',  
516 *Nucleic Acids Research*, 44(13), pp. 6398–6408. doi: 10.1093/nar/gkw325.

517 Licht, K., Kapoor, U., *et al.* (2019) 'A high resolution A-to-I editing map in the mouse identifies  
518 editing events controlled by pre-mRNA splicing', *Genome R*, 29, pp. 1453–1463. doi:  
519 10.1101/gr.242636.118.Freely.

520 Licht, K., Hartl, M., *et al.* (2019) 'NAR Breakthrough Article Inosine induces context-dependent  
 521 recoding and translational stalling', *Nucleic Acids Research*. Oxford University Press, 47(1), pp. 3–  
 522 14. doi: 10.1093/nar/gky1163.

523 Liu, X. *et al.* (2017) 'Comparative analysis of alternative polyadenylation in *S. cerevisiae* and *S.*  
 524 *pombe*', pp. 1685–1695. doi: 10.1101/gr.222331.117.27.

525 Maas, S. and Rich, A. (2000) 'Changing genetic information through RNA editing', *Bioessays*.  
 526 Wiley Online Library, 22(9), pp. 790–802.

527 Macdonald, C. C. (2019) 'Tissue-specific mechanisms of alternative polyadenylation : Testis , brain  
 528 , and beyond ( 2018 update )', (August 2018), pp. 1–11. doi: 10.1002/wrna.1526.

529 Marioni, J. C. *et al.* (2008) 'RNA-seq: An assessment of technical reproducibility and comparison  
 530 with gene expression arrays', *Genome Research*, 18(9), pp. 1509–1517. doi:  
 531 10.1101/gr.079558.108.

532 Martínez, P. *et al.* (2019) 'Multiple evidences suggest sox2 as the main driver of a young and  
 533 complex sex determining ZW/ZZ system in turbot (*Scophthalmus maximus*)', *bioRxiv*, p. 834556.  
 534 doi: <https://doi.org/10.1101/834556>.

535 Mayba, O. *et al.* (2014) 'MBASED: Allele-specific expression detection in cancer tissues and cell  
 536 lines', *Genome Biology*, 15(8), pp. 1–21. doi: 10.1186/s13059-014-0405-3.

537 Mayr, C. (2017) 'Regulation by 3'-Untranslated Regions', *Annual Review of Genetics*, 51, pp. 171–  
 538 194.

539 Mercer, T. R. *et al.* (2012) 'letters Targeted RNA sequencing reveals the deep complexity of the  
 540 human transcriptome', *Nature Biotechnology*. Nature Publishing Group, 30(1), pp. 99–107. doi:  
 541 10.1038/nbt.2024.

542 Meyer, K. D. and Jaffrey, S. R. (2014) 'The dynamic epitranscriptome : N<sup>6</sup>-methyladenosine and  
 543 gene expression control', *Molecular Cell Biology*. Nature Publishing Group, 15(May), pp. 313–326.  
 544 doi: 10.1038/nrm3785.

545 Miao, Z. *et al.* (2018) 'ASElux: An ultra-fast and accurate allelic reads counter', *Bioinformatics*,  
 546 34(8), pp. 1313–1320. doi: 10.1093/bioinformatics/btx762.

547 Monsu, M. and Comin, M. (2021) 'Fast alignment of reads to a variation graph with application to  
 548 SNP detection', *Journal for Integrative Bioinformatics*, 18(4), pp. 1–9.

549 Nachtergaele, S. and He, C. (2017) 'The emerging biology of RNA post-transcriptional  
 550 modifications', *RNA biology*. Taylor & Francis, 14(2), pp. 156–163. doi:  
 551 10.1080/15476286.2016.1267096.

552 Ohta, T. (1995) 'Synonymous and Nonsynonymous Substitutions in Mammalian Genes and the  
 553 Nearly Neutral Theory', *Journal of Molecular Evolution*, 40, pp. 56–63.

554 Ozsolak, F. and Milos, P. M. (2011) 'RNA sequencing: Advances, challenges and opportunities',



555 *Nature Reviews Genetics*. Nature Publishing Group, 12(2), pp. 87–98. doi: 10.1038/nrg2934.  
 556 Pandey, R. V. *et al.* (2013) ‘Allelic imbalance metre (Allim), a new tool for measuring allele-  
 557 specific gene expression with RNA-seq data’, *Molecular Ecology Resources*, 13(4), pp. 740–745.  
 558 doi: 10.1111/1755-0998.12110.  
 559 Panousis, N. I. *et al.* (2014) ‘Allelic mapping bias in RNA-sequencing is not a major confounder in  
 560 eQTL studies’, *Genome Biology*, 15(15), pp. 1–8.  
 561 Piccirillo, C. A. *et al.* (2014) ‘Translational control of immune responses: from transcripts to  
 562 translatoemes’, *Nature Immunology*, 15(6), pp. 503–511. doi: 10.1038/ni.2891.  
 563 Pickrell, J. K. *et al.* (2010) ‘Understanding mechanisms underlying human gene expression  
 564 variation with RNA sequencing’, *Nature*. Nature Publishing Group, 464(7289), pp. 768–772. doi:  
 565 10.1038/nature08872.  
 566 Poplin, R. *et al.* (2017) ‘Scaling accurate genetic variant discovery to tens of thousands of samples’,  
 567 *bioRxiv*, p. 201178. doi: 10.1101/201178.  
 568 Rak, R. *et al.* (2021) ‘Dynamic changes in tRNA modifications and abundance during T cell  
 569 activation’, *Cell Biology*, 118(42), pp. 1–12. doi: 10.1073/pnas.2106556118.  
 570 Robinson, J. T. *et al.* (2017) ‘Variant review with the integrative genomics viewer’, *Cancer*  
 571 *Research*, 77(21), pp. e31–e34. doi: 10.1158/0008-5472.CAN-17-0337.  
 572 Robledo, D. *et al.* (2019) ‘Discovery and functional annotation of quantitative trait loci affecting  
 573 resistance to Sea lice in Atlantic salmon’, *Frontiers in Genetics*, 10(FEB), pp. 1–10. doi:  
 574 10.3389/fgene.2019.00056.  
 575 Root, L., Campo, A., Macniven, L., Con, P., Cnaani, A. and Kùltz, D. (2021) ‘A data-independent  
 576 acquisition ( DIA ) assay library for quantitation of environmental effects on the kidney proteome  
 577 of *Oreochromis niloticus*’, *Molecular Ecology Resources*. doi:  
 578 10.22541/au.160553713.37893872/v1.  
 579 Root, L., Campo, A., Macniven, L., Con, P., Cnaani, A. and Kùltz, D. (2021) ‘Nonlinear effects of  
 580 environmental salinity on the gill transcriptome versus proteome of *Oreochromis niloticus* modulate  
 581 epithelial cell turnover’, *Genomics*. Elsevier Inc., 113(5), pp. 3235–3249. doi:  
 582 10.1016/j.ygeno.2021.07.016.  
 583 Rozowsky, J. *et al.* (2011) ‘AlleleSeq: Analysis of allele-specific expression and binding in a  
 584 network framework’, *Molecular Systems Biology*. Nature Publishing Group, 7(522), pp. 1–15. doi:  
 585 10.1038/msb.2011.54.  
 586 Salavati, M. *et al.* (2019) ‘Elimination of Reference Mapping Bias Reveals Robust Immune Related  
 587 Allele-Specific Expression in Crossbred Sheep’, *Frontiers in Genetics*, 10(September), pp. 1–16.  
 588 doi: 10.3389/fgene.2019.00863.  
 589 Sanfilippo, P., Wen, J. and Lai, E. C. (2017) ‘Landscape and evolution of tissue-specific alternative

polyadenylation across *Drosophila* species', *Genome Biology*. *Genome Biology*, 229(18), pp. 1–22.  
doi: 10.1186/s13059-017-1358-0.

Schwartz, S. *et al.* (2014) 'Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA', *Cell*, 159(September), pp. 148–162.

Sheng, K. *et al.* (2017) 'Effective detection of variation in single-cell transcriptomes using MATQ-seq', *Nature Methods*, 14(3), pp. 267–274. doi: 10.1038/nmeth.4145.

Shikanai, T. (2015) 'Biochimica et Biophysica Acta RNA editing in plants : Machinery and flexibility of site recognition ☆', *BBA - Bioenergetics*. Elsevier B.V., 1847(9), pp. 779–785. doi: 10.1016/j.bbabi.2014.12.010.

Simpson, L. and Thiemann, O. H. (1995) 'Sense from Nonsense : RNA Editing in Mitochondria of Kinetoplastid Protozoa and Slime Molds', *Cell*, 81, pp. 837–840.

Skelly, D. A. *et al.* (2011) 'A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data', *Genome Research*, 21(10), pp. 1728–1737. doi: 10.1101/gr.119784.110.

Song, C., Yi, C. and He, C. (2012) 'Mapping recently identified nucleotide variants in the genome and transcriptome', *Nature Biotechnology*. Nature Publishing Group, 30(11), pp. 1107–1117. doi: 10.1038/nbt.2398.

Stevenson, K. R., Coolon, J. D. and Wittkopp, P. J. (2013) 'Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome', *BMC Genomics*, 14(1), pp. 1–13. doi: 10.1186/1471-2164-14-536.

Stuart, K. (1991) 'RNA editing in mitochondrial mRNA of trypanosomatids', *Trends in Biochemical Sciences*, 16, pp. 68–72. doi: [https://doi.org/10.1016/0968-0004\(91\)90027-S](https://doi.org/10.1016/0968-0004(91)90027-S).

Sun, W. *et al.* (2016) 'RMBase : a resource for decoding the landscape of RNA modifications from high-throughput sequencing data', *Nucleic Acids Research*, 44(October 2015), pp. 259–265. doi: 10.1093/nar/gkv1036.

Takenaka, M. *et al.* (2014) 'RNA editing in plant mitochondria —Connecting RNA target sequences and acting proteins', *Mitochondrion*, 19, pp. 191–197. doi: <https://doi.org/10.1016/j.mito.2014.04.005>.

Trapnell, C. *et al.* (2011) 'Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms', *Nature Biotechnology*, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.Transcript.

Vijaya Satya, R., Zavaljevski, N. and Reifman, J. (2012) 'A new strategy to reduce allelic bias in RNA-Seq readmapping', *Nucleic Acids Research*, 40(16), pp. 1–9. doi: 10.1093/nar/gks425.

Wang, J. (2017) 'The computer program structure for assigning individuals to populations: easy to use but easier to misuse', *Molecular Ecology Resources*, 17(5), pp. 981–990. doi: 10.1111/1755-

0998.12650.

Wang, K., Li, M. and Hakonarson, H. (2010) ‘ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data’, *Nucleic Acids Research*, 38(16), pp. 1–7. doi: 10.1093/nar/gkq603.

Weng, L. *et al.* (2016) ‘Poly ( A ) code analyses reveal key determinants for tissue-specific mRNA alternative polyadenylation’, *Bioinformatics*, 22, pp. 813–821. doi: 10.1261/rna.055681.115.4.

Wolf, T. *et al.* (2020) ‘Dynamics in protein translation sustaining T cell preparedness’, *Nature Immunology*. Springer US, 21(August), pp. 927–940. doi: 10.1038/s41590-020-0714-5.

Xin, H. *et al.* (2013) ‘Accelerating read mapping with FastHASH’, *BMC Genomics*, 14(Suppl 1), pp. 1–13. doi: 10.1186/1471-2164-14-S1-S13.

Yang, S. H. *et al.* (2017) ‘The ultrastructural characterization of mitochondria-rich cells as a response to variations in salinity in two types of teleostean pseudobranch: milkfish (*Chanos chanos*) and Mozambique tilapia (*Oreochromis mossambicus*)’, *Journal of Morphology*, 278(3), pp. 390–402. doi: 10.1002/jmor.20650.

Yuan, S. and Qin, Z. (2012) ‘Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression’, *Proceedings - 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012*, pp. 718–724. doi: 10.1109/BIBMW.2012.6470225.

Zhan, S., Griswold, C. and Lukens, L. (2021) ‘Zea mays RNA-seq estimated transcript abundances are strongly affected by read mapping bias’, *BMC Genomics*. BMC Genomics, 22:285, pp. 1–12. doi: <https://doi.org/10.1186/s12864-021-07577-3>.

Zhang, F. *et al.* (2020) ‘Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits’, *BMC Genomics*. BMC Genomics, 21(1), pp. 1–22. doi: 10.1186/s12864-019-6362-1.

Zviran, A. *et al.* (2019) ‘Deterministic Somatic Cell Reprogramming Involves Continuous Transcriptional Changes Governed by Resource Deterministic Somatic Cell Reprogramming Involves Continuous Transcriptional Changes Governed by Myc and Epigenetic-Driven Modules’, *Cell stem cell*. Elsevier Inc., 24(February), pp. 328–341. doi: 10.1016/j.stem.2018.11.014.

656 **Tables:**

Test description	Test number	SNPs in ASE	Genes containing ASE	DEGs with SNPs in ASE
Gills and kidney in fresh water	1	1731	963	899
Gills and kidney in salty water	2	2311	1254	1153
Gills in fresh and salty water	3	236	193	15
Kidney in fresh and salty water	4	1126	715	0

657

658 **Table 1:** Number of SNPs obtained in each test after applying the pipeline. Chi-sq test on the allele  
659 frequency data for each individual. Five individuals in each group, n=10 for each test, p<0.05.

660

Group	T-statistic	P-value
GF	17,40	1,57E-66
GS	8,62	8,14E-18
KF	9,47	3,65E-21
KS	4,66	3,22E-06

661

662 **Table 2:** Statistics and p-value of each T-student test. The test was performed on the abundance of  
663 the allele frequencies called against the reference genome and called against the pseudogenome, n=  
664 3,740 SNPs. One test was performed in each experimental group: GF gills fresh water, GS gills  
665 salty water, KF kidney fresh water, KS kidney salty water.

666

SNP type	Test 1	Test 2	Test 3	Test 4
UTR3	1064	1379	122	613
CDS	474	611	63	347
ncRNA_exonic	50	106	23	58
UTR5	45	66	7	37
Downstream	38	51	2	24
Intergenic	26	40	11	20
Intronic	20	28	2	13
upstream\3bdownstream	7	12	1	4
Upstream	5	9	4	7
ncRNA_intronic	2	8	1	2
synonymous_SNV	432	548	54	316
nonsynonymous_SNV	35	49	7	23
Ratio synonymous/non-synonymous	12,34	11,18	7,71	13,74
nonframeshift_deletion	3	4	1	3
nonframeshift_insertion	1	2	0	1
frameshift_insertion	1	4	1	1
Unknown	1	1	0	1
frameshift_deletion	1	3	0	2
UTR5\3bUTR3	0	1	0	1

667

668 **Table 3:** Type of SNPs in ASE found for each test, namely: test 1 gills and kidney in fresh water,  
669 test 2 gills and kidney in salty water, test 3 gills in fresh and salty water, test 4 kidney in fresh and  
670 salty water; Chi-sq test,  $p < 0.05$

671

672 **Figures**

673 **Figure 1:** Schema representing the key steps of the pipeline. The RNA is extracted from the  
674 experimental samples and sequenced for obtaining of the fastq files. These files are trimmed and  
675 after quality filters are mapped to the reference genome for a first SNP calling. The biallelic variant  
676 sites obtained in this first call are then used for the creation of two pseudogenomes. The fastq files  
677 are then mapped twice, one to each pseudogenome, and the SNP call is performed also twice. The  
678 resulting variant call files are then submitted to home scripts for the merging and averaging of the  
679 allele depths. The pipeline is developed in snakemake and the scripts are submitted to GitHub.

680

681 **Figure 2: A.** Calculation of the average allele depth after the calling to the two pseudogenomes.  
682 The first genotype of the allele found in the first sample for a site is considered as reference and it is  
683 marked in blue. The next genotype found for the same site in the allele is set as alternative and it is  
684 marked in orange. Only biallelic sites are considered in this pipeline. The figure illustrates the  
685 possibilities of homozygosis and heterozygosis, as well as the no expression after a sample is called  
686 to pseudogenome 1 (PSG1) or pseudogenome 2 (PSG2). **B.** Assignment of the reference and  
687 alternative alleles after homozygosis in one pseudogenome, or different order of the alleles in each  
688 pseudogenome. The first sample is established as a model for the reference and alternative alleles.  
689 The next samples reorganize their position following the model of the first sample. If a site is found  
690 expressed in a sample different than the first model, the position of the alleles is set as the next  
691 sample where there is expression (symbol \*).

692

693 **Figure 3:** T-student test on each experimental group. The test was performed on the abundance of  
694 the allele frequencies called against the reference genome and called against the pseudogenome,  $n =$   
695 3,740 SNPs. One test was performed in each experimental group.

696

697 **Figure 4:** Classification of the ASE SNPs by the type predicted from the coordinates as set in the  
698 annotation. On the right the classification of the non-coding section and on the left the classification  
699 of the coding section. Tests are as follows: : **test 1** gills and kidney in fresh water, **test 2** gills and  
700 kidney in salty water, **test 3** gills in fresh and salty water, **test 4** kidney in fresh and salty water.

701

702 **Figure 5:** Venn graph illustrating the common ASE SNPs for each test. Absolute values of SNPs.

703

704 **Figure 6:** Heatmap including the allele frequencies of the total SNPs found in each experimental  
705 group. The linkage group is described on the left.

706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735

**Figure 7:** GO functions and percentage of the ASE SNPs found in each test.

**Figure 8:** Manhattan plot generated with the ASE SNPs found in each tests.

**Supplementary 1:** Table including ASE SNPs contained in genes that are differentially expressed between gills and kidney in fresh water (test 1). Coordinates of chromosome and position, as well as CHI statistic, product description and GO function associated.

**Supplementary 2:** Table including ASE SNPs contained in genes that are differentially expressed between gills and kidney in salty water (test 2). Coordinates of chromosome and position, as well as CHI statistic, product description and GO function associated.

**Supplementary 3:** Table including ASE SNPs contained in genes that are differentially expressed between gills in fresh and salty water (test 3). Coordinates of chromosome and position, as well as CHI statistic, product description and GO function associated.

**Supplementary 4:** Table including ASE SNPs contained in genes that are differentially expressed between kidney in fresh and salty water (test 4). Coordinates of chromosome and position, as well as CHI statistic, product description and GO function associated.

**Supplementary 5:** Figure including the distribution of the allele frequencies on the experimental groups. **A**, gills in fresh water called against the pseudogenomes. **B**, gills in fresh water called against the reference genome. **C**, gills in salty water called against the pseudogenomes. **D**, gills in salty water called against the reference genome. **E**, kidney in fresh water called against the pseudogenomes. **F**, kidney in fresh water called against the reference genome. **G**, kidney in salty water called against the pseudogenomes. **H**, kidney in salty water called against the reference genomes.

**Supplementary 6:** Table including the differentially expressed genes between gills and kidney in fresh water (test 1) that present ASE SNPs for the same test.

**Supplementary 7:** Table including the differentially expressed genes between gills and kidney in fresh water (test 2) that present ASE SNPs for the same test.

**Supplementary 8:** Table including the differentially expressed genes between gills and kidney in fresh water (test 3) that present ASE SNPs for the same test.